Verkettung von Daten: Record Linkage am Beispiel des Philadelphia Social History Project

Hershberg, Theodore; Burstein, Alan; Dockhorn, Robert

Theodore Hershberg, Alan Burstein, Robert Dockhorn

Verkettung von Daten. Record Linkage am Beispiel des Philadelphia Social History Project*

Wie bei einem großen Teil der *neuen* Sozialgeschichte stand auch bei unseren Forschungen die Beschäftigung mit den Lebenserfahrungen von *gewöhnlichen* Menschen im Mittelpunkt. Dadurch, daß uns für diese Zeit die Manuskripte der Volkszählung zur Verfügung stehen, worin jeder Einwohner der USA mit Namen und anderen wertvollen sozioökonomischen und demographischen Merkmalen aufgeführt wird, sind die Jahre von 1850 bis 1880 zum Brennpunkt für die Rekonstruktion von individuellen Lebensläufen geworden, obwohl derlei Unternehmungen keineswegs nur auf diese Zeit beschränkt blieben. Das Verfahren, womit diese Längsschnitts-Daten gesammelt werden, ist als *Datenverkettung (Record-Linkage)* oder als das Zusammenführen von Angaben zu einem bestimmten historischen Individuum aus unterschiedlichen Quellen bekannt geworden[1].

Eine solche Verkettung von Daten ist keineswegs auf Einzelpersonen beschränkt; sie kann ebenso Nachforschungen über Familien oder Unternehmen und Produktionsbetriebe einschließen. Ebensowenig ist sie auf das Manuskript der Volkszählung als Datenquelle beschränkt; wie dem Genealogen seit langem bekannt, gibt es vielmehr eine Vielzahl von Quellen, dazu gehören Stadt-Adreßbücher, Steuerunterlagen, Personenstandsregister, Nachlaßregister und ähnliches. Schließlich läßt sich die Verkettung von Daten nicht nur im Längsschnitt oder vertikal über die Zeit hin durchführen, Daten können auch im Querschnitt oder horizontal zwischen verschiedenen Quellen aus ungefähr der gleichen Zeit verkettet werden. Bei der Datenver-

* Original erschienen unter dem Titel *Record Linkage*, in: Historical Methods Newsletter, 9, März-Juni 1976, S. 137–163; übersetzt und abgedruckt mit freundlicher Genehmigung der Verfasser und der Herausgeber von *HMN*. Der Herausgeber dankt Frau Angelika Schweikhart (Berlin) für die Übersetzung.

1 Die beste allgemeine Einführung in dieses Thema ist: E.A. Wrigley (Hrsg.), *Identifying People in the Past*, London 1973. Neben einigen ausgezeichneten Artikeln enthält diese Sammlung einen Überblick über die Literatur zur Datenverkettung von Ian Winchester.

kettung im *Philadephia Social History Project (PSHP)* sind viele verschiedene Inhalte, Dokumente und Richtungen enthalten[2].

Mit wenigen Ausnahmen wurde die Verkettung von Daten auf der Ebene von Fallstudien durchgeführt, wo den Individuen longitudinal durch die Manuskripte der Volkszählung oder Adreßbücher nachgespürt wurde. Beim *PSHP* zum Beispiel können Daten nicht verkettet werden, wenn das jeweilige Individuum nicht wieder in den Manuskripten der nächsten Volkszählung für Philadelphia County erscheint. Dies kann durch Tod, einen Fehler oder Abwanderung verursacht sein. Um ein Individuum aufzunehmen, das an einen Ort außerhalb der Stadtgrenzen gezogen ist, müßte man die Manuskripte der Volkszählung der übrigen Nation durchsuchen. Diese Aufgabe aber läge weit außerhalb des Rahmens sowohl unserer Forschung als auch von Fallstudien zu lokalen Gemeinden überhaupt[3]. Wir können daher zum Geschick der Abwanderer nach ihrem Wegzug aus Philadelphia keine Aussagen machen. Wir können darüber hinaus noch nicht einmal feststellen, ob die Erfahrungen derer, die weggezogen sind, den Erfahrungen derer, die geblieben sind, entsprechen. Dies sind die wesentlichen Schwächen einer auf der Ebene der Fallstudie durchgeführten Verkettung von Daten. Dagegen läßt sich aber bei jeder Volkszählung ein sozioökonomisches Profil sowohl von Abwanderern als auch von Zugezogenen erstellen. Denen, die auf die Beschränkungen der Fallstudie verweisen, läßt sich also einfach antworten, daß unsere Forschungen einen spezifisch lokalen Schwerpunkt haben. Wir befassen uns mit der Frage, wer kam, wer zog weg, und was geschah mit denen, die in der zweiten Hälfte des neunzehnten Jahrhunderts in der zweitgrößten Stadt der Nation blieben.

Obgleich das Interesse der Historiker an sozialer Mobilität die Anwendung der Datenverkettung bemerkenswert hat ansteigen lassen, wurde der dabei tatsächlich angewandten Technik bis in allerjüngste Zeit nur wenig systematische Aufmerksamkeit gewidmet. Wie Ian Winchester bemerkt, wurden „die Verkettungsverfahren, die z. B. Griffen in seiner Untersuchung über die Arbeiter von Poughkeepsie oder Gutman in seinen Arbeiten über Paterson, New Jersey oder Thernstrom in seinen Untersuchungen zur beruflichen Mobilität in Newburyport oder Knights in seiner Arbeit über Zu- und Abwanderung, Ortsfestigkeit und innerstädtische Mobilität in Boston anwendet, nicht genauer spezifiziert"[4]. Daß in den ersten Untersuchungen

2 Obwohl unsere Bemühungen bis heute hauptsächlich auf das Verketten von Einzelpersonen nach den Manuskripten der alle zehn Jahre stattfindenden Volkszählung gerichtet waren, haben wir in jedem Zensusjahr auch die Produktionsbetriebe mit den Unternehmen verkettet, um die für unsere Arbeit so wichtigen Angaben zu den Anschriften von letzteren auf erstere zu übertragen.

3 Von zwei Forschern gibt es methodische Anregungen, wie diese Beschränkungen der Fallstudie überspielt werden können. Siehe: Charles Stephenson, *Tracing Those Who Left: Mobility Studies and the Soundex Indexes to the U.S.Census,* in: Journal of Urban History, 1, November 1974, S. 73–85. Siehe auch die Arbeiten von Peter Knights.

4 Siehe: Ian Winchester, *The Linkage of Historical Records by Man and Computer: Techniques and Problems,* in: Journal of Interdisciplinary History, 1, Herbst 1970, S. 107–125. Dieser bahnbrechende Artikel ging aus der Forschung über Hamilton (Ontario), die von Michael

der Verkettung von Daten keine Aufmerksamkeit geschenkt wurde, ist verständlich. Wir erinnern uns, das Problem, wie verkettet wird, nicht einmal bedacht zu haben, als wir Thernstroms Arbeit über Newburyport zum ersten Mal lasen. Aber auch nach Winchesters Beobachtungen von 1970 wird nur in den allerwenigsten veröffentlichten Arbeiten das Verfahren, nach dem die Verkettung der Daten erfolgte, beschrieben. *Historical Methods Newsletter* veröffentlicht vor kurzem zwei wichtige Aufsätze über die Verkettung von Daten[5], aber darin ist primär die technische Seite des Verfahrens in den Vordergrund gerückt; erforderlich aber wären sorgfältige Untersuchungen über die Art und die Folgen der spezifischen *biases*, die mit der Anwendung verschiedener Methoden der Datenverkettung entstehen.

Wir sind uns nicht im klaren darüber, warum die Verkettung von Daten von seiten der Historiker so selten systematisch behandelt worden ist. Wo das Interesse bei der Analyse von Veränderungen in den Lebensläufen von Individuen in der Zeit lag, war wohl der Wunsch „voranzukommen" schuld daran, daß der Datenverkettung und den sehr gravierenden Folgen, die sich aus der Anwendung verschiedener Techniken ergeben, nur flüchtige Überlegungen gewidmet wurden. Und endlich sind die Historiker – als Berufsgruppe – nicht gerade dafür bekannt, daß sie methodologische Studien hochschätzen oder sich sehr dafür interessieren. Aus welchen Gründen immer eine kritische Untersuchung der Datenverkettung unterblieben ist, es bleibt beunruhigend, denn es sollte Klarheit darüber herrschen, daß – wie intelligent auch immer der Forscher sein mag – mangelhaft verkettete Daten zu unzutreffenden Analysen führen. Den Schlußfolgerungen aus Untersuchungen, in denen eine beträchtliche Anzahl von Personen derart fehlerhaft verknüpft oder ein *bias* übersehen wurde, ist zu mißtrauen.

Zum Teil rührt das Problem daher, daß zwischen wichtigen, aber verschiedenen analytischen Zielen der Datenverkettung nicht unterschieden wurde, nämlich der Untersuchung von Ortsfestigkeit (*persistence*) oder Zu- und Wegzug der Bevölkerung, und der Untersuchung von sozialer, beruflicher, geographischer und ökonomischer Mobilität. Der Anteil an einer gegebenen Bevölkerung, der über einen Zeitraum hin ortsfest blieb, wurde mit einer Vielzahl von signifikanten historischen Überlegungen gekoppelt. Wenn, wie bei einer Fallstudie üblich, die Ortsfestigkeit als abhängige Variable genommen wird, verwenden wir Variablen wie Alter, Familienstand, Größe der Familie und ihre Zusammensetzung, Beruf, Vermögen und ethnische Zugehörigkeit zur Erklärung, warum in einer bestimmten Umgebung (Stadt,

Katz geplant wurde und heute als *Canadian Social History Project* bekannt ist, hervor. Wir sind Katz und Winchester für ihre Arbeiten zur Datenverkettung tief verpflichtet. Auf den von ihnen gelegten Fundamenten haben wir ein maschinelles und allgemeines Datenverkettungsprogramm mit mehr Möglichkeiten entwickelt, als es ihr Programm zu Beginn der Forschungen über Hamilton hatte. Heute benutzen Michael Katz und sein Forscherteam die neueste Version unseres Datenverkettungsprogramms.

5 Siehe Michael Katz und John Tiller, *Record Linkage for Everyman: A Semi-Automated Process*, in: Historical Methods Newsletter, 5, September 1972, S. 144–150; Dennis Kelly, *Linking Nineteenth Century Manuscript Census Records: A Computer Strategy*, in: Historical Methods Newsletter, 7, März 1974, S. 72–82.

Stadtgebiet, Gebietseinheit) manche wegziehen und andere bleiben. Wenn für verschiedene Städte zu verschiedenen Zeitpunkten in der Vergangenheit Quoten für die Ortsfestigkeit bekannt wären, könnte untersucht werden, auf welche Weise Unterschiede zwischen den Städten, wie etwa Größe, Lage, Geschichte, Alter, Wirtschaft, Wachstumsraten, Bevölkerungsstruktur, Zugang zu billigen Massentransportmitteln usw. diese Quoten beeinflussen. Wenn Ortsfestigkeit als unabhängige Variable genommen wird, könnte das Ausmaß des Bevölkerungswandels zur Erklärung anderer signifikanter sozialer Phänomene verwendet werden. So sind zum Beispiel niedere Ortsfestigkeitsraten zur Erklärung für fehlendes Klassenbewußtsein, begrenzte Militanz der Arbeiterklasse, langsam wachsende Organisationen der Arbeiter, Instabilität der Gemeinschaft, eine Vielfalt sozialpathologischer Erscheinungen, fortdauernde Herrschaft einer kleinen lokalen Elite und für eine stabile Sozialstruktur herangezogen worden[6].

Es ist nicht nötig, hier alle die vielen Hypothesen aufzuführen, die mit einer historischen Untersuchung der sozialen Mobilität verbunden sind; unsere Absicht ist vielmehr darzulegen, warum zur Bestimmung der Ortsfestigkeitsraten eine andere Methode der Datenverkettung als zur Bestimmung der Raten und Muster der sozialen Mobilität angewandt werden sollte. Als Beispiel sollen die Manuskripte der Volkszählung dienen. Die Listen der Volkszählung geben für jede Person sechs Grundvariablen an: Name, Alter, Geburtsort, Beruf, Adresse und Vermögen. Da durch eine Untersuchung der Ortsfestigkeit herausgefunden werden soll, wie viele Personen, die zum Zeitpunkt T_1 ansässig waren auch an T_2 ansässig waren, können und sollen zur Identifikation alle sechs Variablen verwendet werden. Durch eine Untersuchung der sozialen Mobilität soll herausgefunden werden, welche Veränderungen bei Beruf, Wohnort und Vermögen einer Person stattgefunden haben, daher sollen nur die Variablen, die nicht auf ihre Veränderungen in der Zeit hin untersucht werden (in diesem Fall Name, Alter, Geburtsort), zur Identifikation herangezogen werden. Würden auch die anderen drei Variablen in das Identifikationsverfahren einbezogen, würde man damit zugleich einen *bias* bei den verketteten Dateien einführen. Je mehr Personen auf Grund der Tatsache, daß sie denselben Beruf haben, unter derselben Adresse leben und dasselbe Vermögen besitzen, von einer Volkszählung zur nächsten als identisch behandelt werden, desto größer ist der *bias* der Dateien zur Stabilität hin. Ein solches Vorgehen brächte unvermeidlich Dateien mit potentiell schwerwiegenden *biases* hervor.

Mit anderen Worten: weil das Verfahren der Datenverkettung sowohl der Untersuchung der Ortsfestigkeit als auch der sozialen Mobilität zugrundeliegt, muß es unter Berücksichtigung der verschiedenen Anwendungsmöglichkeiten durchgeführt

6 Siehe: Michael Katz, *The People of Hamilton, Canada-West,* Kapitel 3, *Transiency and Social Mobility;* Howard Chudacoff, *Mobile Americans, 1972;* Th. Hershberg, Rezension zu Chudacoff, in: International Migration Review, 1976; Stephan Thernstrom und Peter Knights, *Men in Motion: Some Data and Speculations About Urban Population Mobility in Nineteenth-Century America,* in: Journal of Interdisciplinary History, 1, Herbst 1975, S. 7–36; Peter Knights, *The Plain People of Boston,* New York 1973.

werden. Man sollte deshalb darauf achten, daß die Dateien, in denen beim Verkettungsprozeß alle vorhandenen Variablen verwendet werden (Untersuchungen zur Ortsfestigkeit), und die, in denen nur die Variablen verwendet werden, die nicht auf Veränderungen in der Zeit untersucht werden (Untersuchungen zur sozialen Mobilität), streng getrennt bleiben. Da auch eine Fragestellung, die auf beiderlei Arten verkettete Daten einbezieht (wie oft sind Personen mit unverändertem Wohnsitz beruflich mobil?), vollkommen legitim ist, vermerkt man am besten auf jedem maschinenlesbaren Datensatz die Art der Verkettung.

Die oben beschriebene Unterscheidung ist nur schwer aufrecht zu erhalten, wenn die Verkettung manuell durchgeführt wird und dabei alle Variablen zu jeder Person deutlich im Blickfeld liegen. Der Versuchung, eine zweideutige Verbindung zu klären, indem man überprüft, ob Beruf, Adresse oder Vermögen zu beiden Zeitpunkten gleichgeblieben sind, läßt sich nur schwer widerstehen. Und doch würden gerade durch dies Verfahren *biases* eingeführt. Zu den Vorteilen maschineller Verkettungstechniken gehört auch, daß solche Versuchungen durch die Verwendung von Computerprogrammen, die nur die für den besonderen Zweck der Verkettung relevanten Variablen heranziehen, ausgeschaltet werden.

Das Thema *bias* wird weiter erhellt, wenn man die zu bearbeitenden Quellen über die Einwohner unter zwei Gesichtspunkten prüft: wie vollständig wird die Bevölkerung erfaßt und welche Angaben werden zu jeder Person in der Quelle gemacht. Zuerst soll die Frage der Vollständigkeit behandelt werden. In Stadt-Adreßbüchern sind mehr Einwohner erfaßt als in Steuerlisten[7], aber weniger als in den Manuskripten der Volkszählung. Jede Quelle ist auf ziemlich eindeutige Weise voreingenommen: die Steuerlisten gegen die Besitzlosen, die Adreßbücher gegen Nicht-Haushaltungsvorstände, Durchreisende, Arme, Schwarze und Einwanderer, die Volkszählungslisten gegen eine Teilmenge von ihnen. Diese Überlegungen werden besonders wichtig, wenn eine Verkettung von der einen zur anderen Quelle versucht wird. Wird von einer umfassenderen Quelle wie den Manuskripten der Volkszählung zu einer weniger umfassenden wie den Adreßbüchern verkettet, so führt das zu niedereren Ortsfestigkeitsraten als beim Verketten von Volkszählung zu Volkszählung oder von Adreßbuch zu Adreßbuch, weil in den Adreßbüchern weniger Personen erfaßt sind als bei der Volkszählung.

Sodann ist zu bedenken, welche verschiedenen Arten der Information in den Datenquellen vorliegen. Von den Sozialhistorikern werden als Dokumente aus dem neunzehnten Jahrhundert am häufigsten die Manuskripte der Volkszählungen und Stadt-Adreßbücher benutzt. In den letzteren werden nur Name, Adresse und Beruf aufgeführt, während in jenen diese Variablen zusätzlich zu Alter, Geburtsort und Vermögen und noch anderen analytisch brauchbaren Informationen enthalten sind. Doch bei den meisten Untersuchungen geschieht die Datenverkettung von Adreßbuch zu Adreßbuch und von Volkszählungsunterlagen zu Adreßbuch. Nur bei wenigen, vor allem wenn es sich um hohe Bevölkerungszahlen handelt, wird von Volks-

[7] Peter Knights, *City Directories as Aids to Ante Bellum Urban Studies: A Research Note*, in: Historical Methods Newsletter, 2, September 1969.

zählung zu Volkszählung verkettet, und dies aus einem sehr einfachen Grund: Adreßbücher haben zwei Vorteile: sie sind gedruckt und garantieren damit Lesbarkeit, weit wichtiger aber noch, sie sind alphabetisch geordnet; die Manuskripte der Volkszählung sind es – trotz der vergleichsweise reichhaltigen Details – nicht. Wie bei der Verwendung von Bezirksstatistiken an Stelle von Daten für kleinere und analytisch brauchbareren Gebietseinheiten durch die Sozialhistoriker, bestimmt auch hier eher die Bequemlichkeit als die innere Qualität der Dokumente den Gebrauch der Adreßbücher. Diese Bequemlichkeit rührt sicherlich nicht unbedingt aus einer Fehleinschätzung der höherwertigen Information der Volkszählungsmanuskripte zur Identifikation und Analyse her und auch nicht vom Widerstreben des Forschers, seine Zeit mit der Arbeit an nicht alphabetisierten Manuskripten zuzubringen (obwohl beides wahrscheinlich vorkommt), oft rührt sie vielmehr von den untragbar hohen Kosten her, die für die Aufarbeitung der handschriftlich vorliegenden Daten der Volkszählung in maschinenlesbare Form entstünden.

Dennoch erhebt sich die Frage, welcher „Preis" gezahlt werden muß, wenn Datensätze von Adreßbuch zu Adreßbuch oder von Volkszählungsunterlagen zu Adreßbuch verkettet werden[8]. Auf welcher Grundlage werden bei einem solchen Verfahren Personen zu T_2 identifiziert? Wie wird technisch *John Jones, Bote, 423 S. Front* zum Zeitpunkt T_1 unter den 54 Personen mit den Namen *John Jones* zum Zeitpunkt T_2 identifiziert, ohne daß sein Beruf oder seine Adresse bei der Identifikation herangezogen wird? Wäre das Forschungsziel nur die Untersuchung der Ortsfestigkeit, so wäre ein solches Verfahren vollkommen gerechtfertigt, aber in einer Untersuchung zur Mobilität würde die Verwendung dieser Angabe, wie wir gesehen haben, die Einführung eines *bias* bedeuten. Um diesen *bias* zu vermeiden, haben manche Forscher den Zugang über *seltene* Namen gewählt. Personen mit häufigen Nachnamen, wie *Jones, O'Connor* oder *Schmidt* würden in einer Stichprobe von Individuen, die in der Zeit verkettet werden sollen, nicht aufgenommen.

Der Ausschluß von Personen mit häufigen Namen aus der verketteten Datei bringt eine Reihe wichtiger Fragen mit sich. Haben Personen mit seltenen Namen zu einem bestimmten Zeitpunkt und über eine Zeit hin andere sozioökonomische Merkmale als Personen mit häufigen Namen? In seiner Antwort auf die von Peter Knights und Richard Alcorn in einer der letzten Nummern von *Historical Methods Newsletter* zu diesem Punkt vorgebrachte Kritik schrieb Stephan Thernstrom, daß „das Schlüsselproblem der Interpretation – ob nämlich verfolgbare Individuen andere Mobilitätserfahrungen haben als nicht verfolgbare – nur durch Studien mit maschinenlesbaren Massendaten wie das *PSHP* gelöst werden kann, das nicht auf Informationen, wie sie in den Adreßbüchern und anderen alphabetischen Listen zur Verfügung stehen, aufbaut, um Dateien zu verketten"[9].

8 Die Verkettung von Volkszählungslisten zu Adreßbuch ist das wünschenswertere von beiden, da die Stichprobe, die mit den Volkszählungslisten gemacht werden kann, den Vorteil der reichhaltigen analytischen Information hat, die das Adreßbuch nicht bietet; doch kann eine Stichprobe noch relativ leicht mit dem alphabetischen Adreßbuch von T_2 verkettet werden.

9 Siehe: Richard S. Alcorn und Peter R. Knights, *Most Common Bostonians: A Critique of*

Die von Knights und Alcorn vorgebrachte Kritik hat zwei unterschiedliche Anliegen. Das erste beschäftigt sich mit der Einführung eines ethnischen *bias*, der sich aus der Eliminierung der Personen mit häufigen Familiennamen ergibt. Letztendlich werden wir diese Frage klären können, doch der derzeitige Gebrauch von ethnisch gesonderten Dateien läßt eine Antwort nicht zu. Für das zweite Problem, ob (innerhalb einer ethnischen Gruppe) Personen mit seltenen Namen sich von denen mit häufigen Namen unterscheiden, haben wir klare Antworten. Obwohl auch die Verkettung der Daten aus der Volkszählung einen eindeutigen *bias* zugunsten von Personen mit seltenen Namen hat, setzen uns doch die zur Identifikation besser geeigneten Angaben in den Manuskripten der Volkszählung in die Lage, auch einige Personen mit häufigen Namen zu verknüpfen. Daher sind wir in gewissem Umfang imstande „festzustellen, ob in den herkömmlichen Verfahren ein systematischer *bias* eingebaut ist"[10]. Auch wenn Forscher, die den Zugang über seltene Namen und über Quellen von der Art der Adreßbücher wählen, durch die die Ergebnisse unserer unten beschriebenen empirischen Studie in gewisser Weise bestätigt werden, haben wir doch ernste Bedenken gegenüber Datensätzen, die an Hand von Adreßbüchern verkettet werden, wenn die analytischen Ziele die Verwendung der Berufs- und Adressenangabe im Identifikationsprozeß eigentlich ausschließen. Während sich die empirischen und interpretatorischen Konsequenzen einer vorwiegend mit Adreßbüchern und weniger mit Volkszählungsmanuskripten ausgeführten Datenverkettung erst noch zeigen müssen, bringt uns unsere Erfahrung dazu, die Verkettung mit Hilfe solcher Quellen zu empfehlen, bei denen die Identifikation auf der Grundlage von Übereinstimmungen beim Alter, Geburtsort und vielleicht bei den Angaben über Frau und Familie möglich ist und weniger auf einer bloßen Namensübereinstimmung beruht.

Die einzelnen Schritte bei der Datenverkettung

Beim *PSHP* war das Ziel der Datenverkettung die Erstellung von Längsschnittdateien über Einzelpersonen aus den Manuskripten der Volkszählung, die alle zehn Jahre durchgeführt worden ist. Die bislang vervollständigten verketteten Dateien schließen männliche Erwachsene irischer und deutscher Abstammung in jedem dieser drei Jahrzehnte zwischen 1850 und 1880 ein und wurden eigens für die Untersuchung der Determinanten und Muster der geographischen, beruflichen und ökonomischen Mobilität erstellt[11]. Theoretisch ist das Verfahren, von einzelnen Dateien

Stephan Thernstrom's 'The Other Bostonians', 1880–1970, in: Historical Methods Newsletter, 8, Juni 1975, S. 98–114; das Zitat stammt aus Thernstrom: *Rejoinder to Alcorn and Knights,* Ibid., S. 117.

10 Ibid.

11 Wie im vorhergehenden Abschnitt dargestellt wurden die Dateien zur Untersuchung der Mobilität und nicht der Ortsfestigkeitsraten aufgebaut. Zusätzliche Informationen über die Erstellung von verketteten Dateien wie auch Beispiele für eine Anwendung auf die Untersuchung der Mobilität, finden sich bei: Alan N. Burstein, *Residential Distribution and Mobility of Irish*

zu verketteten überzugehen, nicht kompliziert. Felligi und Sunter haben ein ausgeklügeltes mathematisches Modell entwickelt, das den Prozeß der Datenverkettung ausführlich beschreibt[12]. Es ist nicht schwierig, aus ihrer Darstellung ein leicht verständliches Modell, das den gesamten Prozeß abdeckt, zu synthetisieren, und eine solche Synthese stellt das Modell dar, das wir benutzt haben. Der Prozeß der Datenverkettung, wie er durch dies Modell beschrieben wird, hat drei Begriffe zur Voraussetzung: Vergleichsraum, Vergleichsfunktion und Entscheidungsregel. Der *Vergleichsraum* stammt von den beiden zu verkettenden Dateien und besteht aus all den Paaren, die man erhält, wenn aus jeder Anfangsdatei ein Individuum genommen wird. Die *Vergleichsfunktion* ist die Weise, in der die Glieder eines jeden Paares verglichen werden, und das Ergebnis eines jeden solchen Vergleichs unterliegt einer *Entscheidungsregel*, die festlegt, ob ein Paar als *verkettet*, *nicht-verkettet* oder als *unbestimmt* gilt.

Während das Modell der Datenverkettung theoretisch einfach ist, erweist sich seine Anwendung auf die Manuskripte der Volkszählungen des neunzehnten Jahrhunderts zur Erstellung von tatsächlich verketteten Dateien als komplex. Wesen und Qualität der Daten lassen eine ganze Reihe von Problemen entstehen. Inkonsistenz bei der Schreibweise von Namen, besonders bei Familiennamen, ist die Regel. Auch Fehler bei der Übertragung der handgeschriebenen Manuskripte in maschinenlesbare Form sind unvermeidlich. In den Manuskripten kommen nicht nur häufig falsche Schreibweisen vor, es finden sich sogar direkte Namensänderungen. Dies war besonders beim Aufbau der verketteten Dateien mit deutschen Namen hinderlich. Die Verwendung amerikanischer Laute für die Schreibweise der deutschen Namen ist ein Problem, das bis zu einem gewissen Grad umgangen werden kann; die direkte Übersetzung von Namen allerdings, durch die ein *Schwarz* von 1850 zu einem *Black* im Jahr 1860 wird, ist ein nicht zu lösendes Problem, es sei denn, die deutsche und die amerikanische Version sind phonetisch ähnlich, wie *Schmidt* und *Smith;* Vornamen, besonders die häufigen, werden seltener falsch geschrieben, obwohl der *John* des einen Jahres in nächsten zu einem *Jonathan* werden kann. Auch hier stellte sich wieder die Amerikanisierung der Vornamen als Problem, das in den deutschen Dateien häufiger vorkam als in den irischen. Bei der Erstellung der verketteten Dateien ist neben dem Namen auch das Alter ein Identifikator von offenkundigem Wert. Die Erwartung jedoch, daß nach einem Jahrzehnt ein verkettetes Individuum in den Manuskripten immer als zehn – auch neun oder elf – Jahre älter erscheinen müßte, sollte aufgegeben werden; falsche Altersangaben sind in den Manuskripten der Volkszählungen an der Tagesordnung[13]. Die Probleme sind zwar zahl-

and German Immigrants in Philadelphia, 1850–1880, Ph. D. Diss., University of Pennsylvania 1975.

12 Ivan Feligi und Alan Sunter, *A Theory for Record Linkage,* Journal of the American Statistical Association, 64, Dezember 1969, S. 1183–1210.

13 Dennis Kelly fand bei der Datenverkettung mit den Manuskripten der Volkszählung in Birmingham, New York und Johnstown, Pennsylvania heraus, daß Einzelpaare mit einer Diskrepanz bei der Altersangabe von bis zu acht Jahren, d. h. die im Verlauf eines Jahrzehnts den Unterlagen nach zwischen zwei und achtzehn Jahre älter geworden waren, als mögliche Verket-

reich, aber sie sind zu lösen. Unsere Absicht ist es, solche Lösungen zu erörtern, indem wir darstellen, wie wir das Grundmodell der Datenverkettung auf die Erstellung von verketteten Dateien angewendet haben. Dabei benutzten wir ein Software-Paket, das im Rahmen des *PSHP* entwickelt worden ist[14].

Vergleichsraum: PSHP-Soundex-System

Zur Spezifizierung des Vergleichsraums müssen zwei Eingabedateien vorhanden sein, aus denen dann das Datenverkettungsprogramm Paare bildet. Das Verkettungsprogramm führt zu einer Datei, in der die Eingabedaten für jedes Paar zusammengestellt sind. Daher sollten die in den Eingabedateien gespeicherten Daten alle die Informationen enthalten, die zur Analyse der schließlich verketteten Dateien relevant sind.

Die Eingabedateien, die wir *unverkettete* Dateien nennen, unterscheiden sich in zweifacher Hinsicht von den Dateien, die nur auf den Angaben zu einem einzelnen Individuum in den Zensusmanuskripten basieren. Zum einen fanden wir heraus, daß sich die Ergebnisse der Verkettung erheblich verbesserten, wenn zum Vergleich nicht nur die Informationen herangezogen wurden, die sich in den Zensusunterlagen direkt auf das betreffende (männliche) Individuum beziehen, sondern auch die Informationen über seine Frau. Daher wurden in die unverketteten Dateien auch der Name, das Alter und der Geburtsort der Frau mit aufgenommen; ebenso sollte auch mit anderen zusätzlichen Identifikatoren, die bei dem Vergleich verwendet werden sollen, wie etwa Informationen über die Kinder oder andere Mitglieder des Haushaltes, verfahren werden.

Die andere Modifikation der unverketteten Dateien betrifft das Problem der Namensschreibung. Um den Schwierigkeiten mit falschen Schreibweisen zu begegnen, wurde bei der Datenverkettung im medizinischen oder kommerziellen Bereich vielfach das *Russel Soundex Coding System* verwendet[15]. Danach erhält jeder Name einen Code aus vier Zeichen. Das erste Zeichen ist durch den Anfangsbuchstaben des Namens bestimmt, die übrigen drei Zeichen bestehen aus numerischen Codes für Gruppen von ähnlich klingenden Konsonanten. Binnen- und Endvokale werden

tungen angesehen werden sollten. Siehe: Dennis Kelly, *Linking Nineteenth Century Manuscripts Census Records.* Wir entschieden uns für denselben Spielraum und fanden ihn angemessen, obwohl unser Programm einige richtig verkettete Paare fand, die noch außerhalb dieses Intervalls lagen, sodaß es sicher auch noch weitere gibt, die sowohl wir als auch Kelly verfehlten.

14 Ein Bericht über die Abweichung und anfängliche Implementierung des Programms gibt: Terry Stickel, *A Computerized Nominal Record Linkage System,* M.S. Thesis, Moore School of Electrical Engineering, University of Pennsylvania 1972; Terry Stickel und Rachel Bodle, *Stages in Record Linkage,* unveröffentlichtes Manuskript, Philadelphia Social History Project, University of Pennsylvania 1972. Die derzeitige Version enthält wichtige Verbesserungen bei der Auswahl der Variablen, der Berechnung der Gewichtungen und der Auswahl der Abschneidepunkte, die in die früheren Berichte von Stickel und Bodle nicht aufgenommen waren.

15 Ian Winchester, *The Linkage of Historical Records,* S. 114–117.

ebenso wie die Konsonanten *H* und *W* ignoriert, bestimmte Endkonsonanten gleichfalls; doppelte Konsonanten werden wie einfache kodiert.

Ohne die entsprechenden Veränderungen umginge das *Russel-Soundex-System* viele der in unseren Dateien gefundenen Fehler. Doch hat der Russel-Code zwei schwerwiegende Mängel. Erstens wurde das System entworfen, um mit Eigenheiten der Schreibweise bei angelsächsischen Namen fertig zu werden. Während das System mit nur geringfügigen Änderungen für irische Namen angemessen sein wird, ist es dies für deutsche Namen weit weniger[16]. Um die Besonderheiten bei der Schreibweise von deutschen Namen auszugleichen, wurden daher Veränderungen der Codierung vorgenommen.

Der zweite Mangel, den das Russel-System für diesen Anwendungsbereich hat, liegt darin, daß es für die Lösung von Problemen entworfen wurde, die sich aus den klanglichen Ähnlichkeiten verschiedener Buchstabenkombinationen ergeben. Aber bei der Arbeit mit Dateien, die auf der Transkription handgeschriebener Manuskripte beruhen, müssen nicht nur ähnlich klingende, sondern auch ähnlich aussehende Buchstaben berücksichtigt werden. Daher haben wir zusätzlich auf ein Codierungssystem wie *Viewex* zurückgegriffen, weil es auch die optische Interpretation der Namensschreibung berücksichtigt[17]. Die Codierung, die wir schließlich verwendeten (sie ist in *Tab. 1* zusammengefaßt), enthält Eigenschaften sowohl des *Soundex-* als auch des *Viewex*-Systems. Bei jedem solchen Codierungssystem müssen Vor- und Nachteile abgewogen werden. Ein zu starres Codierungssystem wird eine große Anzahl von äquivalenten Namen nicht als solche erkennen, während ein zu offenes System denselben Code zu vielen verschiedenen Namen zuordnen würde. Die endgültige Grundlage für den Vergleichsraum besteht dann aus den Unterlagen der einzelnen Dateien, denen die Codes für den Vor- und Nachnamen, den Vornamen der Frau sowie ihr Alter und ihr Geburtsort hinzugesetzt sind.

16 Ähnliche Schwierigkeiten sind bei der Anwendung des Russel-Soundex auf andere Sprachen aufgetreten. Siehe z. B.: Jaques Legare, Yolande Lavoie und Hubert Charbonneau, *The Early Canadian Population: Problems in Automatic Record Linkage*, in: The Canadian Historical Review, 53, Dezember 1972; die Autoren befassen sich mit den Unzulänglichkeiten des *Russel-Soundex* für die Codierung von französischen Namen.

17 Winchester, *The Linkage of Historical Records*, S. 117.

Tabelle 1

Regeln für die Codierung nach dem PSHP Soundex-Viewex System

(Die Schritte erfolgen in der unten aufgeführten Reihenfolge)

I Codierung des Familiennamens

A. Behandlung des Präfix

1. VON, VAN wird abgetrennt
2. MC, MAC wird 1
3. O' wird O[1]
4. Anfangs- E, I, O, U wird A
5. GE, GI, GY wird JE, JI, JY
6. CE, CI, CY wird SE, SI, SY
   CHR wird KHR;
   sonstiges CH bleibt CH
   sonstiges C wird zu K
7. KN wird zu NN
8. PH wird zu FH
9. WIE wird zu VIE
   WEI wird zu VEI
10. W wird zu V[2]
11. M wird zu N[2]
12. Y wird zu J[2]
13. Z wird zu S[2]

B. Behandlung des Suffix

1. End- R wird zu N
2. End- SE, CE fallen weg
3. End- S, SS fallen weg
4. STOWN wird zu SAWON[1]
5. MPSON wird zu MASON[1]
6. NSEN wird zu ASEN[1]
   MSON wird zu ASON[1]
7. STEN wird zu SAEN[1]
   STON wird zu SAON[1]
8. NG wird zu N
   ND wird zu N
9. GAN wird AAN[1]
   GEN wird AEN[1]
10. End- TES, TS fallen weg[2]
11. End- TZE, ZE fallen weg[2]
12. End- Z, TE fallen weg[2]

C. Behandlung des Infix

1. CK wird zu C
2. SCH wird zu S
3. DT wird zu T

Fortsetzung Tabelle 1

4. ND wird zu N
5. NG wird zu N
6. LM wird zu M
7. MN wird zu M
8. WIE wird zu VIE
9. WEI wird zu VEI

D. Soundex-Codierung

1. Nach der Behandlung des Präfix wird der erste Buchstabe behalten
2. A, E, I, O, U, W, H werden übergangen
3. B, P, F, V erhalten den Code 1
4. C, G, J, K, S, Q, Z erhalten den Code 2
5. D, T erhalten den Code 3
6. L erhält den Code 4
7. M, N, R erhalten den Code 5
8. Doppelbuchstaben werden wie einzelne behandelt.

II. Codierung des Vornamens und des Namens der Ehefrau

A. Behandlung des Präfix

1. GE, GI, GY wird zu JE, JI, JY
2. CE, CY, CI wird zu SE, SY, SI
3. CHR wird zu KHR
   sonstiges CH bleibt CH
   sonstiges C wird zu K
4. KN wird zu NN
5. PH wird zu FH
6. WIE wird zu VIE
   WEI wird zu VEI

B. Soundex-Codierung

1. Der erste Buchstabe des Namens nach der Behandlung des Praefix wird behalten
2. A, E, I, O, U, W, H werden übergangen
3. B, P, F, V erhalten den Code 1
4. C, G, J, K, X, Q, S, Z erhalten den Code 2
5. D, T erhalten den Code 3
6. L erhält den Code 4
7. M, N, R erhalten den Code 5
8. Doppelbuchstaben werden wie einzelne behandelt
9. Nur die erste Silbe des Namens wird codiert
10. JAMES erhält den Code J700 zur Unterscheidung von JOHN (J500)
11. PAT. . . erhält den Code P700 zur Unterscheidung von PETER (P300)

---

[1] Nicht für die deutsche Datei

[2] Nur für die deutsche Datei

Vergleichsfunktion: Vorwahlvariablen und Binärgewichtung

Betrachtet man einen Vergleichsraum, der auf zwei großen Dateien mit je 25 000 Individuen basiert, dann besteht in diesem Fall der Vergleichsraum aus 625 000 000 Datensatzpaaren. Alle diese Paare einem detaillierten Vergleich nach jedem einzelnen Merkmal zu unterziehen, würde untragbar viel Rechenzeit und Geld kosten. Um die Effizienz des Computers bei diesem Verkettungsverfahren zu erhöhen, sollte die Zahl der Paare, die einem genauen Vergleich unterzogen werden, drastisch reduziert werden. Daher muß die Durchführung des Vergleichs in zwei Phasen unterteilt werden, die Winchester den *Suchschritt* und den *detaillierten Vergleichsschritt* genannt hat[18]. Zweck des Suchschrittes ist, die Anzahl der Paare, die einem detaillierten Vergleich unterzogen werden sollen, auf eine handhabbare Größe zu reduzieren. Der Suchschritt wählt die Paare mit einer oder mehreren übereinstimmenden *Vorwahlvariablen (pocket-variables)* aus, das sind Merkmale, die bei den meisten zutreffend verketteten Paaren, aber nur bei einem kleinen Teil der nicht zusammengehörenden Paare übereinstimmen. Da die Paare nur dann einem detaillierten Vergleich als mögliche verkettete Paare unterzogen werden, wenn die Vorwahlvariablen übereinstimmen, ist klar, daß – sofern es nicht eine Vorwahlvariable gibt, die bei allen zutreffend verketteten Paaren übereinstimmt – die Paare, bei denen die Vorwahlvariable nicht übereinstimmt, bei dieser Methode nicht verkettet werden. Wenn eine Vorwahlvariable nicht genügend diskriminant ist, um eine große Anzahl von Zufallspaaren auszuscheiden, dann bleibt die Zahl der Paare, die einem detaillierten Vergleich unterzogen werden unannehmbar hoch. Daher setzt der Suchschritt notwendig das Abwägen zwischen den Unkosten des detaillierten Vergleichs und dem Anteil aller zutreffenden Verkettungen, die verglichen werden könnten, voraus. Wieviele zusammengehörenden Paare bei dem Suchschritt verloren gehen, wird durch die für den detaillierten Vergleich verfügbare Rechenzeit bestimmt, da der Suchschritt ein relativ billiges Verfahren ist.

Bei der Erstellung unserer eigenen verketteten Dateien wurden die *Soundex*-Codes des Vor- und Familiennamens als Vorwahlvariable benutzt[19]. Hätten wir nur das *Soundex* des Familiennamens verwendet, so wäre zweifellos eine Anzahl wirklich verketteter Paare, bei denen Änderungen der Schreibweise besonders ausgeprägt waren, unberücksichtigt geblieben; gleichwohl war das Familiennamen-*Soundex* die einzige hinreichend diskriminante Variable, um die Größe der Vorwahlgruppe auf ein vernünftiges Maß zu reduzieren. Um den Suchschritt effizienter zu gestalten, wurden die noch nicht verketteten Dateien vom Computer in alphanumerischer Ordnung entsprechend den Vorwahlvariablen sortiert, sodaß die Vorwahl-

18 Ibid., S. 113–114.

19 Kelly, *Linking Nineteenth Century Manuscripts Census Records* verwendete den Vornamen und die letzte Initiale als Vorwahlvariable bei einem halbmaschinellen System. Der Umfang unserer Dateien hätte die Wahl dieser Vorwahlvariablen unvertretbar teuer werden lassen, da die resultierenden Vorwahlgruppen unerträglich groß geworden wären.

gruppen für den Vergleich vom Computer mit Hilfe eines einfachen billigen Programms leicht ausgewählt werden konnten.

Wir benutzten für die Vorwahlvariablen beim Suchschritt und für die Vergleichsvariablen beim detaillierten Vergleichsschritt dasselbe *Soundex-System.* Dies hätte nicht so sein müssen. Damit ein Paar verkettet werden kann, müssen alle Vorwahlvariablen, aber wie später zu sehen sein wird, nicht alle Vergleichsvariablen übereinstimmen. Ein *Soundex-System,* das im detaillierten Vergleichsschritt genauer und diskriminanter als im Suchschritt wäre, würde verkettete Dateien erzeugen, bei denen der Anteil an eingeschlossenen, aber nicht zusammengehörenden Paaren wahrscheinlich genauer als bei unserem reduziert wäre.

Wenn die Anzahl der Paare aus dem Vergleichsraum, die für eine Verkettung in Frage kommen, hinreichend reduziert worden ist, kann der detaillierte Vergleich durchgeführt werden. Der detaillierte Vergleichsschritt versucht das Verfahren zu simulieren, das ein Forscher anwendet, wenn er eine verkettete Datei manuell herstellt. Wie läßt sich die Vorgehensweise bei der manuellen Verkettung charakterisieren? Der Forscher betrachtet bei jedem Paar eine Reihe von Variablen, deren Werte entweder übereinstimmen oder voneinander abweichen. Wenn die Werte übereinstimmen, wächst die Vermutung, daß ein Paar verkettet ist. Wie stark diese Vermutung anwächst, ist abhängig von der jeweils zu vergleichenden Variablen und ihrem Wert. So trägt zum Beispiel die Übereinstimmung bei einem häufig vorkommenden Vornamen weniger zu der Vermutung bei, daß ein Paar verkettet ist, als Übereinstimmung bei einem sehr seltenen Vornamen. Weichen die Werte einer Variablen voneinander ab, dann ist die Vermutung, daß ein Paar dennoch zusammengehört, um so unwahrscheinlicher, je wahrscheinlicher es ist, daß diese Variablen bei einem zutreffend verketteten Paar übereinstimmen würden.

Ein Computerprogramm allerdings erfordert ein genauer definiertes Verfahren als die Bestärkung einer Vermutung, aber daran ist abzulesen, welche Richtung einzuschlagen ist. Die Glieder der Paare müssen im Hinblick auf gewisse Variablen verglichen, und das Maß von Übereinstimmung oder Nichtübereinstimmung durch eine Maßzahl ausgedrückt werden: bei Übereinstimmung sollte ein bestimmter Wert hinzu-, bei Nichtübereinstimmung ein Wert abgerechnet werden. Sodann muß ein Gewichtungssystem eingesetzt werden, das für jede Variable festlegt, welcher Wert der Maßzahl eines Paares zugezählt oder von ihr abgezogen werden soll. Winchester beschreibt verschiedene mögliche Gewichtungssysteme, aber dasjenige, das der Vorstellung von einer sich verstärkenden Vermutung am ehesten entspricht, ist das von Newcombe entwickelte System der *Binärgewichtung*[20]. Binärgewichtungen basieren auf dem Verhältnis zwischen der Wahrscheinlichkeit, daß eine Variable bei einem zutreffend verketteten Paar übereinstimmt, und der Wahrscheinlichkeit, daß eine Variable bei einem zufällig ausgewählten Paar übereinstimmt, und die tatsächliche Gewichtung ist der Logarithmus von diesem Verhältnis zur Basis 2. In dem Maß wie die Wahrscheinlichkeit, daß die Werte einer Variablen in einem verketteten Paar übereinstimmen, relativ zur Wahrscheinlichkeit, daß sie in einem zufällig ausgewähl-

20 Winchester, *The Linkage of Historical Records,* S. 119–122.

ten Paar übereinstimmen, zunimmt, werden die Gewichtungen entsprechend höher. Der Übereinstimmung derselben Variablen können in verschiedenen Paaren unterschiedliche Gewichtungen zugeordnet werden, da Übereinstimmung von Variablen mit einem häufigen Wert zu einer niedrigeren Gewichtung führt als Übereinstimmung mit Werten, die weniger häufig sind.

Die aktuelle Computerformel für die Binärgewichtungen kann am besten aus der mathematischen Ableitung verstanden werden. Wenn L den Fall bezeichnet, daß ein Paar zutreffend verkettet ist, und C_i das Ergebnis des Vergleichs der Variablen i für ein Datenpaar ist, dann kann C_i verschiedene Werte annehmen: die Variable stimmt nicht überein, die Variable stimmt überein, oder die Variable stimmt mit einem besonders häufigen oder mit einem besonders seltenen Wert überein. Die kombinierten Ergebnisse der Vergleiche von n Variablen bei einem Paar können dann so angegeben werden: $\overset{n}{\underset{i}{\Omega}} 1C_i$, , die Schnittlinie der n Vergleiche.

Gesucht wird ein Gewichtungssystem, das $\Pr(L|\overset{n}{\underset{i=1}{\Omega}}C_i)$ darstellt, die Wahrscheinlichkeit, daß ein Paar aufgrund der Gesamtergebnisse der Vergleiche von n Variablen zutreffend verkettet ist. Die konditionale Wahrscheinlichkeit kann entsprechend Bayes Theorem transformiert werden:

$$\Pr(L|\overset{n}{\underset{i=1}{\Omega}}C_i) = \Pr(L) \quad \frac{\Pr(\overset{n}{\underset{i=1}{\Omega}}C_i|L)}{\Pr(\overset{n}{\underset{i=1}{\Omega}}C_i)}$$

In dieser Formel bezeichnet Pr(L) die Wahrscheinlichkeit, daß ein zufällig ausgewähltes Paar zutreffend verkettet ist, und $\Pr(_{i\ =\ 1}C_i)$ die Wahrscheinlichkeit, daß alle beobachteten Ergebnisse der Vergleiche von n Variablen bei einem zufällig ausgewählten Paar auftreten, und $\Pr(_{i\ =\ 1}\ C_i/L)$ die Wahrscheinlichkeit, daß alle beobachteten Ergebnisse bei einem zutreffend verketteten Paar auftreten.

Wenn bei den Vergleichen aller n Variablen die Ergebnisse unabhängig verteilt sind, dann ist

$$\Pr(\overset{n}{\underset{i=1}{\Omega}}C_i) = \Pr(C_1)x\Pr(C_2)x\ldots x\Pr(Cn) = \overset{n}{\underset{i=1}{\pi}}\Pr(C_i),$$

und

$$\Pr(L|\overset{n}{\underset{i=1}{\Omega}}C_i) = \Pr(L) \quad \overset{n}{\underset{i=1}{\pi}} \frac{\Pr(C_i|L)}{\Pr(C_i)}$$

Nimmt man den Logarithmus zur Basis 2 von beiden Seiten der Gleichung

$$\log_2(\Pr(L|\overset{n}{\underset{i=1}{\Omega}}C_i)) = \log_2(\Pr(L)) + \overset{n}{\underset{i=1}{\Sigma}} \log_2(\frac{\Pr(C_i|L)}{\Pr(C_i)}).$$

definiert $\quad \mathrm{Wt}(I) = \log_2 \ (\frac{\Pr(C_i|L)}{\Pr(C_i)})$.

dann ist $\quad \log_2 \ (\Pr L|\overset{n}{\underset{i=1}{\Omega}}C_i)) = \log_2(\Pr(L)) + \overset{n}{\underset{i=1}{\Sigma}} \mathrm{Wt}(i).$

Pr(L) ist bei allen Verkettungsproblemen konstant und kann aus der Gleichung entfernt werden, dann hat man Hanlins Bestätigung: $H(L) + \overset{n}{\underset{i=1}{\Sigma}} \mathrm{Wt}(i)$.

Man beachte, daß $H(L) \neq \log_2(\Pr(L_{|i}^{\Omega} 1C_i))$, da *P(L)*, die Wahrscheinlichkeit, daß ein zufällig ausgewähltes Paar zutreffend verkettet wird, aus der Formel eliminiert ist. Dennoch kann die Wahrscheinlichkeit nicht ohne die vorherige Kenntnis der wirklichen Ortsfestigkeitsrate geschätzt werden. *H(L)* stellt ein Substitut für $\Pr(L_{|i}^{\Omega} 1C_i)$ dar, das gesucht wurde, und die Formel für *Wt(i)* wird zur Berechnung der Binärgewichtung verwendet. Zur Berechnung der Gewichtungen braucht man zwei Werte, die Wahrscheinlichkeit, daß ein solches Ergebnis bei einem zufällig ausgewählten Paar auftritt, und die Wahrscheinlichkeit, daß ein solches Ergebnis bei einem zutreffend verketteten Paar auftritt. Der erste läßt sich direkt aus der Verteilung der Werte in den Eingabedateien errechnen. Der zweite muß geschätzt werden, vorzugsweise aus einer stichprobenweise verketteten Datei, aber bei großen Dateien verändern abweichende Schätzungen das Endergebnis des Verkettungsprozesses nicht wesentlich.

Ein wichtiger Aspekt des Systems der Datenverkettung, wie es vom *PSHP* praktiziert wird, ist die Zuordnung unterschiedlicher Gewichtungen je nach der Häufigkeit der Werte einer Variablen in der Anfangsdatei. Zum Beispiel ist unsere Überzeugung, daß ein Datensatzpaar zutreffend verkettet ist, größer bei Übereinstimmung eines seltenen Vornamens als bei häufiger vorkommenden Vornamen. Das System der Binärgewichtung berücksichtigt diese Vorstellung, denn je seltener ein Name in der Eingabedatei gefunden wird, desto höher ist das Verhältnis der Wahrscheinlichkeit, daß bei einem zutreffend verketteten Paar dieser Name übereinstimmt, geteilt durch die Wahrscheinlichkeit, daß dieser Name bei einem zufällig ausgewählten Paar übereinstimmt, und folglich ist die Binärgewichtung, die einem Paar bei Übereinstimmung dieses Namens zugeordnet wird, umso höher.

In der Praxis ordnen wir den besonders häufigen Werten einer Variablen jeweils eigene Binärgewichtungen zu, während wir allen weniger häufigen Werten dieser Variablen eine einzige Gewichtung zuordnen. Ein Beispiel für die Berechnung dieser Gewichtungen wird in *Tabelle 2* gegeben. In diesem Beispiel sollten zwei Dateien

Tabelle 2

Beispiel für die Binärgewichtung des Vornamens

	Anzahl der Fälle				
Name	Datei 1	Datei 2	Pr(Ci)	Pr (Ci|L)	Gewichtung
James	20	22	.078	.266	1.8
John	17	16	.048	.209	2.1
William	14	14	.034	.177	2.3
Thomas	10	9	.016	.120	2.9
Patrick	8	9	.013	.108	3.1
Richard	2	2			
Owen	1	1			
Felix	1	0	.001	.057	5.5
Bernard	0	1			
Edward	2	1			

mit jeweils 75 Namen verkettet werden, und es waren die Gewichtungen für das Ergebnis des Vergleichs der Vornamen zu berechnen. (In diesem Fall wurde unterstellt, daß die Namen richtig geschrieben und nicht nach dem *Soundex-System* codiert sind.) Um die Gewichtungen berechnen zu können, waren zwei Datenquellen nötig. Erstens muß man die Verteilung der Namen in jeder der beiden Eingabedateien kennen, wie in *Tab. 2* gezeigt. Diese Verteilungen sind erforderlich, damit man bestimmen kann, welche Namen häufig und welche weniger häufig sind. Die zweite, für die Berechnung der Gewichtungen notwendige Datenquelle ist eine Häufigkeitsschätzung der verschiedenen Vergleichsergebnisse in einer Gruppe von verketteten Paaren. Im Idealfall ist eine stichprobenweise verkettete Datei für die Bestimmung der Wahrscheinlichkeit vorhanden. Wie genau diese Wahrscheinlichkeit ohne schon vorhandene große verkettete Datei geschätzt werden kann, ist fraglich, aber auch durch Abweichungen bei der Schätzung wird der Wert der Gewichtungen relativ zueinander nicht wesentlich verändert. Bei diesem Beispiel wurde angenommen, daß die Wahrscheinlichkeit für Übereinstimmung des Vornamens bei einem zutreffend verketteten Paar 0.95 ist[21]. Da für häufige Werte der Variablen besondere Gewichtungen verwendet werden, wie in dem Beispiel für *James*, liegt die relevante Wahrscheinlichkeit nicht in dem Ergebnis, daß die Vornamen in einem zutreffend verketteten Paar übereinstimmen, sondern eher darin, daß die Vornamen übereinstimmen und *James* sind in einem zutreffend verketteten Paar. In einer kleinen, stichprobenweise verketteten Datei sind die Häufigkeiten der einzelnen Werte wahrscheinlich zu gering, als daß man jede dieser Wahrscheinlichkeiten exakt schätzen könnte, besonders wenn Personen mit seltenen Vornamen eher verkettet werden als solche mit häufigen Vornamen.

Nach unserer Vorgehensweise wurde die Wahrscheinlichkeit, daß eine Variable mit einem beliebigen Wert übereinstimmt, mit Hilfe der stichprobenweise verketteten Datei geschätzt, und dann wurde angenommen, daß die Werte einer Variablen wie des Namens in einer Datei, die alle zutreffenden Verkettungen enthielte, im gleichen Verhältnis verteilt wären wie in den Eingabedateien. Zum Beispiel haben nach *Tab. 2* 26,7 % der Individuen in *Datei 1* und 29,3 % der Individuen in *Datei 2* den Vornamen *James*. Wenn es keinen Zusammenhang zwischen tatsächlicher Ortsfestigkeit und Vornamen gibt, dann ist von dem Mittelwert dieser Zahlen, von 28 % aller zutreffend verketteten Paare zu erwarten, daß sie den Vornamen *James* haben. Daher beträgt die Wahrscheinlichkeit, daß ein zutreffend verkettetes Paar in beiden Datensätzen den Vornamen *James* hat 0.28 x 0.95 = 0.266. Es ist zu erwarten, daß in der tatsächlichen Menge aller zutreffenden Verkettungen eine Übereinstimmung des Vornamens mit dem Wert *James* bei etwa 27 % auftritt. Diese Zahl würde in der verketteten Datei, die mit Binärgewichtungen aufgebaut wurde, kleiner, da die Übereinstimmung des Vornamens *James* eine geringere Gewichtung erhielte als die Übereinstimmung bei einem selteneren Vornamen.

21 Dies wird als die Wahrscheinlichkeit genommen, daß der Vorname in der Menge aller richtig verketteten Paare übereinstimmt. Wenn der Vorname als Vorwahlvariable verwendet wird, wie beim Aufbau unserer verketteten Datei geschehen, dann stimmen notwendig alle vom Programm hergestellten Verkettungen beim Vornamen überein.

Die Binärgewichtungsformel auf die Teilwahrscheinlichkeiten anzuwenden ist einfach. Für die Übereinstimmung eines Vornamens mit dem Wert *James* wäre die Binärgewichtung:

$$\log_2 \frac{\frac{(20+22)/2}{75} \times 0.95}{\frac{20 \times 22}{75^2}} = 1.8.$$

Analog werden die Binärgewichtungen für alle häufigeren Werte berechnet. Die Trennlinie zwischen häufigeren und weniger häufigen Werten festzulegen, ist ebenfalls nicht schwierig. In unserem Beispiel ist zwischen *Patrick* und *Richard* deutlich eine Lücke zu erkennen. In der Praxis gibt es eine solche Lücke nicht immer, aber in dem Maß, in dem in großen Dateien die Häufigkeiten abnehmen, wachsen die Gewichtungen für die einzelnen Werte bis nahe an das durch die Dateigröße gegebene Maximum, und ein geeigneter Abschneidepunkt *(cutoff point)* zwischen *üblichen* und *weniger üblichen* Werten kann anhand einer einfachen Durchsicht festgelegt werden. So lange sehr häufige Werte als üblich behandelt werden, hat die genaue Lage des Punktes, von dem an Werte als weniger üblich behandelt werden sollen, nur geringe Bedeutung, da einzelne Gewichtungen für besonders seltene Werte sich der Gewichtung nähern, die allen weniger üblichen Werten gemeinsam ist. Die Gewichtung für die weniger üblichen Werte in dem Beispiel ist:

$$\log_2 \frac{\frac{(2+2 + 2+1 + 1+1 + 1+0 + 0+1)/2}{75} \times 0.95}{\frac{2 \times 2}{75^2} + \frac{2 \times 1}{75^2} + \frac{1 \times 1}{75^2} + \frac{1 \times 0}{75^2} + \frac{0 \times 1}{75^2}} = 5.5$$

Die Gewichtung für Nichtübereinstimmung einer Variablen wird nach der Binärgewichtungsformel berechnet und ist der Logarithmus zur Basis 2 von der Wahrscheinlichkeit, daß die Variable bei einem zutreffend verketteten Paar nicht übereinstimmt, geteilt durch die Wahrscheinlichkeit, daß die Variable in einem zufällig ausgewählten Paar nicht übereinstimmt. Da für jede beliebige Variable, die bei der Datenverkettung verwendet wird, Nichtübereinstimmung bei einem zufällig ausgewählten Paar wahrscheinlicher ist als Nichtübereinstimmung in einem verketteten Paar, ist dieser Bruch kleiner als eins und führt zu einer negativen Gewichtung. In unserem Beispiel ist die Gewichtung für Nichtübereinstimmung

$$\log_2 \frac{1 - 0.95}{1 - \left(\frac{20 \times 22}{75^2} + \frac{17 \times 16}{75^2} + \frac{14 \times 14}{75^2} + \frac{10 \times 9}{75^2} + \frac{8 \times 9}{75^2} + \frac{2 \times 2}{75^2} + \frac{1 \times 0}{75^2} + \frac{0 \times 1}{75^2} + \frac{2 \times 1}{75^2} + \frac{1 \times 1}{75^2}\right)} = -4.0.$$

Bei der Berechnung der Gewichtungen für das *Soundex* des Nachnamens fanden wir zu viele übliche *Soundex*-Codes für Nachnamen, als daß unser Programm sie ohne furchtbare Unkosten hätte handhaben können. Um unser System auch auf

Nachnamen anwenden zu können, ordneten wir die *Soundex*-Codes der Nachnamen in 12 Häufigkeitsgruppen und setzten jedem Datensatz einen Code für die Häufigkeit hinzu. Die Menge aller *Soundex*-Codes der Nachnamen wurde dann so unterteilt, daß die Intervalle zwischen den Gewichtungen, die dem Häufigkeitscode zugeordnet waren, annähernd gleich waren. Bei der Durchführung des Vergleichs wurde der Übereinstimmung des Nachnamen*soundex* die Gewichtung Null zugeordnet, während der Übereinstimmung des Häufigkeitscodes, die notwendig einem übereinstimmenden Nachnamen*soundex* folgen muß, eine eigene Gewichtung zugeordnet wurde. Es ist offenkundig, daß sich diese Methode nur anwenden läßt, wenn der Nachname eine Vorwahlvariable ist, denn sonst könnte der Häufigkeitscode leicht übereinstimmen, auch wenn der Code für das Nachnamen*soundex* dies nicht tut.

Die Gewichtungen für den Altersvergleich werden ähnlich berechnet, aber hier ist eher das Intervall zwischen den Altersangaben interessant und nicht die besonderen Werte für das Alter. Darauf wurden Gewichtungen berechnet, und zwar für Personen, die genau zehn Jahre älter geworden waren, für solche, die zwischen neun und elf Jahren älter geworden waren und für solche, die zwischen acht und zwölf Jahren älter geworden waren, bis hin zu denen, die zwischen zwei und achtzehn Jahren älter geworden waren, alle Altersdifferenzen außerhalb dieser Kategorien erhielten eine negative Gewichtung. Die zur Berechnung der Gewichtungen verwendeten Fälle sollten inklusiv sein, d. h. die Gewichtung sollte eher für den Fall errechnet werden, daß ein Individuum mindestens acht und höchstens zwölf Jahre älter ist, und nicht für den Fall, daß ein Individuum genau acht oder zwölf Jahre älter ist. Die Gewichtung, die dann beim Vergleichsschritt tatsächlich zugeordnet wird, ist diejenige für die höchst gewichtete Kategorie, in die ein Paar fällt.

Wegen der algorithmischen Form der Vergleichsfunktion als der Summe der Binärgewichtungen ist es endlich nötig, die bei dem Vergleich tatsächlich verwendeten Variablen zu spezifizieren. Wir haben schon gezeigt, welche Gefahren mit der Verwendung von Variablen verbunden sind, die dann in der eigentlichen Analyse, der die verketteten Dateien unterzogen werden, abhängige Variable sind. Durch die Verwendung der Binärgewichtungen wird auf zwei andere Kriterien verwiesen, denen die Vergleichsvariablen genügen sollten. Erstens muß die Gewichtung, die eine Variable zur Maßzahl des Vergleichs hinzusetzt, die marginalen Kosten hinreichend rechtfertigen, die jede im Durchlauf des Verkettungsprogramms eingesetzte Variable verursacht. Ganz besonders sollte die Wahrscheinlichkeit, daß die Variable in einem zufällig ausgewählten Paar übereinstimmt, verglichen mit der Wahrscheinlichkeit, daß sie in einem verketteten Paar übereinstimmt, gering sein. So ist zum Beispiel der Geburtsort der Frau eine Variable, die beim Vergleichsschritt verwendet werden könnte; doch da die meisten Frauen von irischen Einwanderern selbst in Irland geboren sind, ist die Wahrscheinlichkeit, daß der Geburtsort der Frau in einem zufällig ausgewählten Paar übereinstimmt, hoch. Folglich wäre die Gewichtung für die Übereinstimmung beim Geburtsort der Frau zu gering, als daß ihre Einbeziehung als Variable beim Vergleichsschritt gerechtfertigt wäre.

Das zweite Kriterium für die Auswahl der Variablen, und das ist durch die mathematische Ableitung der Gewichtungen bestimmt, ist, daß die Vergleichsergeb-

nisse verschiedener Variablen unabhängig voneinander sein müssen. In den ersten Versuchen mit dem Verkettungsprogramm wurde das Alter der Frau als Vergleichsvariable verwendet. Da aber die Ehefrau eines Mannes ihm sehr wahrscheinlich im Alter nahesteht, hat ein Datenpaar mit angemessener Abweichung im Alter nach einem Jahrzehnt, sofern in beiden Jahren eine Frau aufgeführt ist, sehr wahrscheinlich auch eine Abweichung in entsprechendem Umfang beim Alter der Frau, unabhängig davon, ob das Datenpaar ein verkettetes Paar darstellt oder nicht. Die Maßzahl, die sich aus dem detaillierten Vergleichsschritt ergibt, ist das Maß für die *Konfidenz*, daß das Paar zusammengehört. Wenn jedoch das Kriterium der Unabhängigkeit verletzt wird, beeinträchtigt die Interaktion der Variablen ernsthaft die Interpretation der Maßzahl, und nachdem das Alter der Frau aus späteren Verkettungsversuchen ausgeschlossen worden war, wurde ein besseres Verhältnis zwischen der Konfidenz, daß das Paar zutreffend verkettet ist, und der durch das Paar erreichten Maßzahl erzielt. Die Erstellung der verketteten Dateien irischer Einwanderer wurde dann auf dem Vergleich von vier Variablen aufgebaut: *Soundex* des Vornamens, *Soundex* des Nachnamens, Alter und dazu das *Soundex* des Namens der Frau. Da von 1860 an für viele deutsche Einwanderer das jeweilige Geburtsland angegeben wurde, wurde bei der Erstellung der deutschen verketteten Dateien für die beiden letzten Jahrzehnte zusätzlich zu den anderen vier Variablen ein Vergleich des Geburtsortes eingesetzt, was zu erheblich genaueren und umfassenderen verketteten Dateien führte.

Daher haben unsere verketteten Dateien einen starken *bias* zugunsten von verheirateten Männern. Die Einbeziehung der Informationen über die Frauen vergrößerte die Richtigkeit der Verkettungen und erlaubte uns sehr viel mehr Verkettungen vorzunehmen, als sonst möglich gewesen wäre. In den verketteten Dateien aber sind unverhältnismäßig wenig ledige Männer enthalten, aber doch hinreichend viele, um die Analyse in weitem Rahmen zu ermöglichen, ungeachtet dessen, daß sie mit größerer Wahrscheinlichkeit Abwanderer sind. Der Familienstand ist natürlich bei jedem Datensatz als Variable mitaufgenommen.

Die Entscheidungsregel: Festsetzung des Abschneidepunktes

Der letzte Schritt des Verkettungsverfahrens ist die Spezifizierung der Entscheidungsregel. Bei welcher durch die Vergleichsfunktion bestimmten Maßzahl wird das Paar der Menge der verketteten Paare, der Menge der nicht verketteten Paare oder der Menge der unbestimmten zugeordnet? Wenn die Variablen richtig ausgewählt sind, dann ist die Wahrscheinlichkeit, daß das Paar zutreffend verkettet ist, um so größer, je höher die Maßzahl, obwohl die direkte Berechnung dieser Wahrscheinlichkeit wegen der Eigenart der Gewichtung nicht möglich ist. Besonders interessant sind drei Maßzahlen. Erstens gibt es, theoretisch wenigstens, eine Maßzahl, genannt X_1, unterhalb derer alle Paare Nicht-Verkettungen sind. Es sollte ebenfalls eine Maßzahl geben, X_2, oberhalb derer alle Paare zutreffend verkettet sind. Paare, deren Maßzahl zwischen X_1 und X_2 liegt, fallen in die Gruppe der unbestimmten. Das

Problem bei der Wahl einer Entscheidungsregel liegt nun darin, einen *Abschneidepunkt (cutoff-point)*, genannt X_3, zu finden, der irgendwo zwischen X_1 und X_2 liegt und über dem nur eine noch tolerierbare Anzahl von Nichtverkettungen der verketteten Datei zugeordnet ist. Will man aber eine Datei, die nur aus richtigen Verkettungen besteht, dann nimmt man X_2 als Abschneidepunkt. Wenn es andererseits wichtiger ist, möglichst viele verkettete Paare zu erfassen als nicht-verkettete Paare auszuschließen, kann X_1 als Abschneidepunkt verwendet werden. Bei der Ableitung unserer verketteten Dateien lag uns mehr an der Minimierung der Anzahl der nicht-verketteten Paare in den Dateien als an der Maximierung der Anzahl der gefundenen zutreffend verketteten Paare. Gleichwohl wird eine Untersuchung mit dem Ziel, zum Beispiel die Eigenart der wirklich Ortsfesten zu bestimmen, besser einen niedrigeren Abschneidepunkt nehmen, um möglichst viele Ortsfeste zu verketten, auch um den Preis, daß in den verketteten Dateien eine große Anzahl nicht zutreffend verketteter Paare enthalten ist.

Während die Berechnung der Gewichtungen von der relativen Häufigkeit der Werte in den Eingabedateien abhängt, beruht die aktuelle Festlegung des Abschneidepunktes auf der absoluten Anzahl der Fälle in einer Eingabedatei. Im Einzelfall wird der Abschneidepunkt so festgelegt, daß die erwartete Anzahl von Individuen in einer Datei mit Merkmalen, die eine Verkettung mit einem Individuum in einer anderen Datei, das dieselben Merkmale hat, ermöglichen, angemessen geringer als eins ist. Auf Grund dieser Überlegungen kann man den Abschneidepunkt relativ einfach berechnen. Die zu erwartende Anzahl von zufälligen Verkettungen, die bei jeder beliebigen Kombination der Werte der Vergleichsvariablen auftreten kann und daher auch bei jeder Maßzahl für Verkettung, läßt sich aus der Häufigkeit der Werte in den Eingabedateien schätzen. Vorausgesetzt, die Ergebnisse aus dem Vergleich verschiedener Variablen sind unabhängig voneinander, wird jede Kombination von Vergleichsergebnissen, die eine höhere Maßzahl erzielt, die zu erwartenden Zufallshäufigkeiten in den verketteten Dateien verringern. Wenn das Kriterium der Unabhängigkeit ernsthaft verletzt ist, wird kein Abschneidepunkt befriedigende Ergebnisse liefern, da die Annahmen, die der Berechnung der Gewichtungen zugrundeliegen, nicht länger gültig wären.

Welche Ansprüche an die Genauigkeit einer verketteten Datei gestellt werden, hängt von ihrer analytischen Anwendung ab. Man sollte beachten, daß zwei Arten von Ungenauigkeit auftreten können. Erstens können in einer Datei Paare enthalten sein, die nicht zutreffend verkettet sind; zweitens können zusammengehörende Paare aus der Datei ausgeschlossen sein. Offenkundig wird durch die Minimierung des einen Fehlertyps der andere häufiger. Die Balance zwischen den zwei auftretenden Fehlertypen wird durch die Auswahl des Abschneidepunktes festgelegt. Wenn der Abschneidepunkt herabgesetzt wird, wächst der Anteil der inkorrekten Verkettungen in der Datei wie auch der Anteil aller eingeschlossenen zutreffenden Verkettungen.

Der Grad der erwünschten Genauigkeit muß nicht schon vor dem Aufbau einer verketteten Datei exakt festgelegt werden. In der Praxis empfiehlt es sich, den Abschneidepunkt absichtlich niedrig anzusetzen, besonders in den frühen Läufen des

Systems, denn dann erhält man eine verkettete Datei, die für eine lohnende Auswertung der Effektivität des Gesamtsystems hinreichend heterogen ist. Die aktuelle Maßzahl der Verkettung sollte auf jedem Datensatz mitaufgeführt werden, damit in späteren Analysen die Paare mit der höchsten Wahrscheinlichkeit, daß sie zutreffend verkettet sind, aus der anfänglich gebildeten Datei herausgezogen und gesondert analysiert werden können.

Gerade die Zufügung der Maßzahl für die Verkettung zu den Datensätzen erlaubt uns zu analysieren, welche Implikationen die verschiedenen Typen der Ungenauigkeit haben. Durch die Erhöhung des Abschneidepunktes und damit der Erhöhung des Anteils an Verkettungen in der Datei, die zutreffen, erhöhten wir in den verketteten Dateien auch den *bias* gegenüber Individuen mit bestimmten Merkmalen, deretwegen sie besonders leicht zu verketten waren. Mit der Verwendung höherer Abschneidepunkte setzte sich die Datei zunehmend aus solchen Individuen zusammen, die verheiratet waren oder ungewöhnliche Namen hatten. Das Problem von *selten vorkommenden* Namen war für manche Forscher besonders interessant, und mit der Verwendung zusätzlicher Identifikationsmerkmale waren wir dann in der Lage, genügend Individuen mit *häufiger vorkommenden* Namen zu finden, um bestimmen zu können, in welchem Ausmaß der *bias* gegenüber selten auftretenden Namen das Ergebnis der Analyse beeinträchtigen könnte.

Selten und häufig vorkommende Namen

Bei allen Arten der Datenverkettung, ob sie nun automatisch oder in einem der vielfältigen Handverfahren vorgenommen werden, läßt sich ein hoher Konfidenzgrad der Verkettung leichter für Individuen mit seltenen Namen als für die Träger von häufig vorkommenden Namen erreichen. Daher enthalten verkettete Dateien nur einen Teil der ortsfesten Bevölkerung und unverhältnismäßig viele Individuen mit selten vorkommenden Namen. Dies wirft die interessante Frage auf, ob Personen mit seltenen Namen irgendwelche besonderen Merkmale gemeinsam haben? Wenn ja, dann kämen durch ihre Überrepräsentation – und die daraus folgende Unterrepräsentation der Personen mit den allerhäufigsten Namen – andere *biases* in die Daten. Auf diese Frage eine definitive Antwort zu finden, ist nicht leicht, denn ob ein *bias* vorhanden ist oder nicht, könnte sehr wohl von der besonderen Eigenart der bei der Verkettung beteiligten Dateien abhängen. Im folgenden werden also die Implikationen der Namenshäufigkeit in einer der ethnischen Dateien des *PSHP* – der irischen Datei für das Jahr 1860 – untersucht, und im Anschluß daran werden die Implikationen für eine multiethnische Bevölkerung erörtert.

1860 teilte sich die große Mehrheit der männlichen Iren nur eine sehr kleine Menge von Vornamen. Von insgesamt 37 303 über 18 Jahre alten Männern in Philadelphia, die Irland als Geburtsort angaben, hießen nicht weniger als 6 897 oder 18,5 % *John*. Andere häufig vorkommende Namen waren *James, William, Patrick, Thomas, Michael, Robert, Daniel* und *Edward*, in dieser Reihenfolge. Insgesamt

68,6 % aller Individuen trugen diese neun Namen. In schroffem Kontrast dazu steht die Verteilung der Nachnamen, wo der häufigste Namen *Kelly* nur bei 463 oder 1,2 % der Männer insgesamt vorkam. Der Häufigkeit nach folgten darauf die Nachnamen *Smith, Dougherty, Murphy, Campbell, McLaughlin, Brown, Wilson, Galagher* und *O'Brien*. Diese zehn Namen zusammen machten aber nur 7,3 % aus. Die große Vielfalt bei den Nachnamen machte für analytische Zwecke eine Zusammenfassung der Namen in Gruppen notwendig, die durch Zuordnung eines *Namenshäufigkeitscodes* von *0* bis *11* erreicht wurde. Der Namenshäufigkeitscode *0* umfaßte die am wenigsten häufigen, Code *11* die häufigsten Namen. Diese Kategorien waren schon für die Zuordnungen von außergewöhnlichen Gewichtungen beim Verkettungsverfahren entwickelt worden (s. o.); es war einfach eine Frage der Bequemlichkeit, dieselben Kategorien erneut anzuwenden.

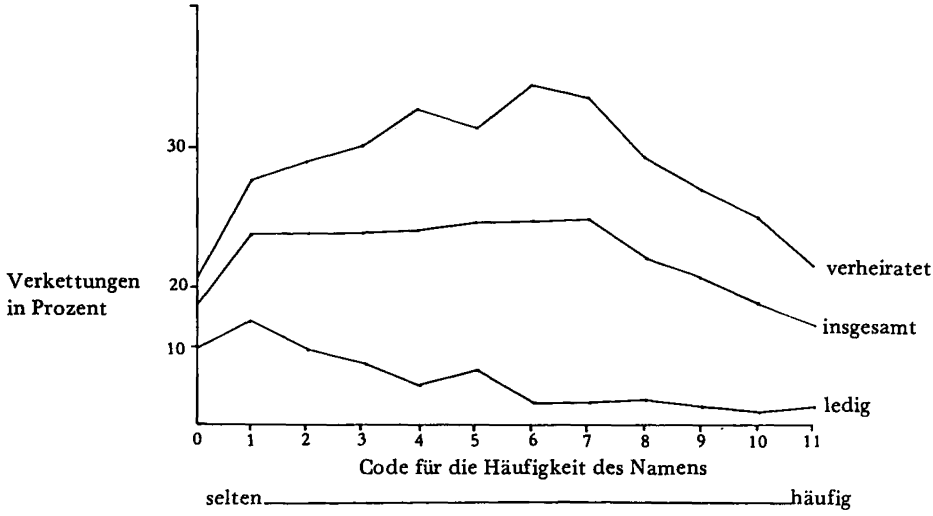
In welchem Ausmaß häufig vorkommende Namen in den verketteten Dateien der männlichen Iren für die Jahre 1860 und 1870 unterrepräsentiert sind, wird aus den *Abb. 1* und *2* ersichtlich. *Abb. 1* zeigt die Ortsfestigkeitsrate nach der Häufigkeit des Nachnamens versus den Code für die Namenshäufigkeit. Vergleicht man das obere und das untere Schaubild miteinander, lassen sich die Veränderungen ablesen, die eintreten, wenn der Abschneidepunkt erhöht wird, um unsichere Verkettungen auszuschließen. Im unteren Schaubild mit dem höheren Abschneidepunkt sind häufig vorkommende Namen deutlich unterrepräsentiert. Dies wirkt sich besonders verheerend bei den männlichen, ledigen Einwohnern aus, aber es betrifft auch die verheirateten Männer, die davon profitieren, daß der Name der Frau als Verkettungsvariable verwendet wird. Man beachte auch die leichte Steigerung auf der äußersten linken Seite, besonders für den Code *0* – für die Namen mit der geringsten Häufigkeit. Das hat seinen Grund wahrscheinlich darin, daß in diese Kategorie eine wesentliche Anzahl falschgeschriebener Namen einbezogen wurde, die nicht verkettet werden konnten, da sie in das falsche *Soundex*-Fach fielen. Daß häufig vorkommende Nachnamen unterrepräsentiert sind, ist der herrschende Trend im unteren Schaubild. Eine ähnliche Unterrepräsentation der häufig vorkommenden Vornamen in der verketteten Datei zeigt *Abb. 2*.

Nachdem erwiesen ist, daß es in der verketteten Datei wirklich einen *bias* bei der Namenshäufigkeit gibt, bleiben noch die sozioökonomischen Unterschiede zwischen Personen mit häufigen und seltenen Namen zu untersuchen. *Tab. 3* enthält eine Kreuztabulation von Alter, Familienstand, Bildung, Besitz und Code für den Beruf *VERT*[22] versus die Kategorien der Nachnamen. Danach lassen sich die weniger häu-

[22] Für ein Codebuch zur Entschlüsselung der Berufscodes, die im Rahmen des PSHP verwendet werden, siehe: Th. Hershberg und Robert Dockhorn, *Occupational Classification*, in: Historical Methods Newsletter, 9, März-Juni 1976, S. 78 ff. Die Hauptklassen sind: VERT 1 *high white-collar* und *professional* Berufe, VERT 2 *low white-collar* und *proprietary* Berufe, VERT 3 gelernte Arbeiter, VERT 4 ungelernte Arbeiter mit spezifierter Tätigkeitsangabe, VERT 5 ungelernte Arbeiter mit unspezifierter Tätigkeit, VERT 6 ungelernte Arbeiter, die weder VERT 4 noch 5 sind, VERT 7 keine Berufsangabe, sondern nur Bezeichnung des Arbeitsplatzes oder des Produktionszweiges, VERT 8 ohne Beruf, VERT 9 nicht zu klassifizieren.

Abb. 1
Prozentsatz der Verkettungen nach der Häufigkeit des Nachnamens
Iren, 1860–1870

(A) Total verkettete Datei (Gewichtung ⩾ 14):



(B) Reduzierte Datei (Gewichtung ⩾ 19):

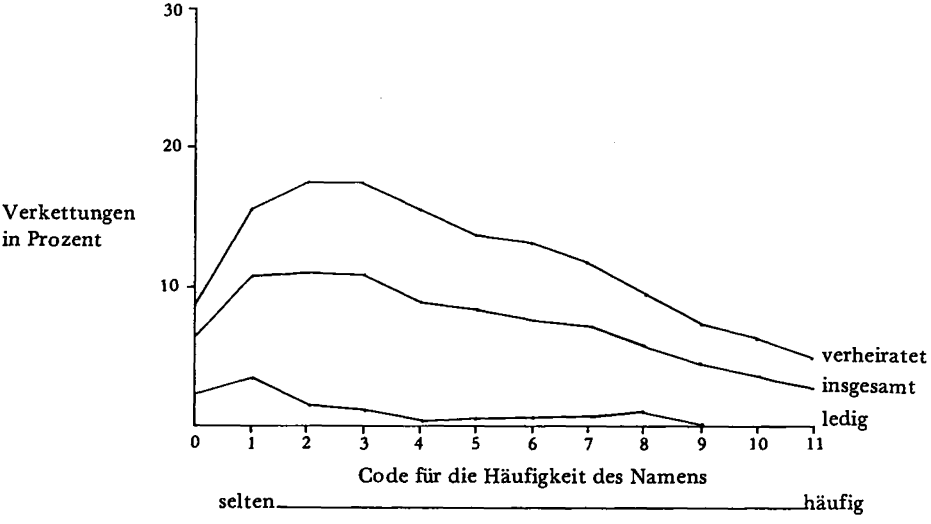
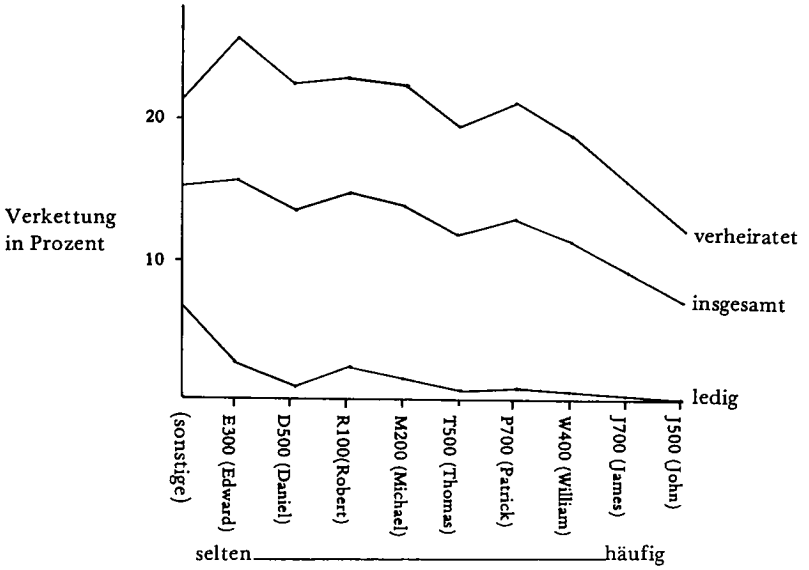
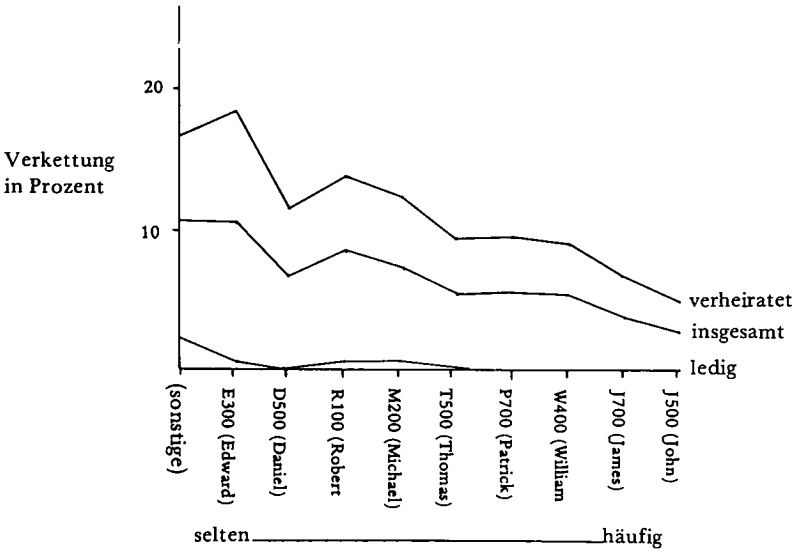


Abb. 2
Prozentsatz der Verkettungen nach dem Vornamen
Iren, 1860–1870

(A) Total verkettete Datei (Gewichtung ⩾ 14):



(B) Reduzierte Datei (Gewichtung ⩾ 19):



figen Namen (Code *0–5*) mit den häufigeren (Code *6–11*), aber ebenso auch die extremen (Code *0* und Code *11*) vergleichen. Tab. 4 gibt dieselben Daten wie *Tab. 3*, aber kreuztabuliert mit den Kategorien der Vornamen. In *Tab. 5* sind dieselben Messungen auf die häufigsten einzelnen Nachnamen und in *Tab. 6* auf die häufigsten einzelnen Vornamen angewandt.

Tabelle 3
Iren 1860
Häufigkeit des Nachnamens nach Alter, Familienstand, Analphabetentum, Eigentum und vertikalem Berufscode

		unter 30 %	verheiratet %	Analphabeten %	mit Grundbesitz %	durchschn. Grundbesitz p. Besitzer	mit Privateigentum %	durchschn. Privateigentum p. Besitzer	Anzahl
geringste Häufigkeit	CODE 0	34.1	57.4	5.4	9.4	$4745	31.0	$727	1579
	CODE 0-5	32.9	58.4	5.8	10.3	4773	32.3	865	12277
	CODE 6-11	33.6	58.4	7.0	10.3	4849	32.9	821	25026
größte Häufigkeit	CODE 11	34.6	58.0	7.1	9.4	4176	31.4	618	4200
Gesamtbevölkerung		33.3	58.4	6.6	10.3	4824	32.7	836	37303

VERT

		0	1	2	3	4	5	6	7	8	9	N
geringste Häufigkeit	CODE 0	9.1	1.4	9.5	35.1	13.3	28.4	1.6	0.9	0.5	0.3	1579
	CODE 0-5	9.2	1.1	10.8	34.3	13.0	28.1	1.8	1.0	0.5	0.2	12277
	CODE 6-11	7.8	0.8	10.8	31.5	13.4	32.2	1.5	1.2	0.6	0.1	25026
größte Häufigkeit	CODE 11	7.5	0.7	10.5	29.9	14.1	33.8	1.6	1.4	0.5	0.2	4200
Gesamtbevölkerung		8.3	0.9	10.8	32.4	13.3	30.9	1.6	1.1	0.6	0.2	37303

Tabelle 4
Iren 1860
Häufigkeit des Vornamens nach Alter, Familienstand, Analphabetentum, Eigentum und vertikalem Berufscode

	unter 30 %	verheiratet %	Analphabeten %	mit Grundbesitz %	durchschn. Grundbesitz p. Besitzer	mit Privateigentum %	durchschn. Privateigentum p. Besitzer	Anzahl
seltenere Namen	32.8	57.5	6.4	10.3	$5206	33.1	$929	11636
häufigere Namen	33.9	58.8	6.7	10.3	4651	32.5	792	25667
Gesamtbevölkerung	33.5	58.4	6.6	10.3	4824	32.7	836	37303

VERT

	0	1	2	3	4	5	6	7	8	9	Anzahl
seltenere Namen	9.7	1.0	11.2	32.3	13.6	28.8	1.3	1.2	0.7	0.2	11636
häufigere Namen	7.6	0.9	10.6	32.5	13.1	31.8	1.7	1.1	0.6	0.1	25667
Gesamtbevölkerung	8.3	0.9	10.8	32.4	13.3	30.9	1.6	1.1	0.6	0.2	37303

Eine Untersuchung der Häufigkeiten in *Tab. 3* und *4* enthüllt zuallererst, daß es zwischen Personen mit häufig und mit selten vorkommenden Namen keine große Disparität gibt. In der Kategorie der Nachnamen zeigt ein Vergleich der Codes *0–5* mit den Codes *6–11* nahezu identische Muster bei Alter, Familienstand und Besitz, ebenso auch starke Ähnlichkeiten bei den meisten Kategorien des vertikalen Berufscodes. Es gibt Unterschiede beim Analphabetentum (5,8 % für selten vorkommende Nachnamen und 7,0 % für häufige), bei *VERT 0* (fehlende Angaben – eine unmittelbare Folge der Codierungsmethoden), bei *VERT 3* (gelernte Arbeiter) und *VERT 5* (ungelernte Arbeiter), aber diese Unterschiede sind nur geringfügig. Diese Konsistenz der Daten durch die Namenskategorien hindurch, sogar bei einem Vergleich der extremen Namenscodes *0* und *11* ist erstaunlich. Dies gilt in gleichem Maß für die Kategorien der Vornamen. Bei einem Vergleich der häufigeren und weniger häufigen Vornamen lassen sich sogar noch größere Ähnlichkeiten als bei den Nachnamen feststellen. Der einzige statistisch signifikante Unterschied beträgt nur 3 % in *VERT 5* (ungelernte Arbeiter).

Dies allgemeine Bild ändert sich aber, wenn man einzelne Namen und nicht Namenskategorien betrachtet (*Tab. 5* und *6*). Hier sind bei allen Angaben größere Abweichungen zu finden. Manche Nachnamen, etwa *Campbell* und *Wilson,* haben eine erheblich günstigere Stellung als etwa *Murphy* und *Gallagher.* Bei Vornamen zeigen sich darüber hinaus sogar noch verblüffendere Unterschiede. Man vergleiche zum Beispiel die vertikale Verteilung der Namen *Robert* und *Patrick.* Während 38,9 % der *Roberts* einen *VERT 3*-Beruf haben, fallen nur 25,5 % der *Patricks* in diese Kategorie. Gegenüber 40,5 % der *Patricks* sind andererseits nur 23,7 % der *Roberts* in *VERT 5.* Diese Unterschiede konstituieren eine bedeutsame Abweichung beim Berufsprofil.

Jeder, der mit den Eigenheiten der Namengebung bei den irischen Einwanderern vertraut ist, kann sich diese Unterschiede erklären. Sie sind auf die Existenz einer subethnischen Teilung von Iren zurückzuführen. Katholiken, die den Hauptteil der irischen Einwanderer in Amerika ausmachten, und Protestanten, von denen viele zuvor aus England oder Schottland nach Irland eingewandert waren, brachten ihre Familiennamen und ihre Eigenheiten der Namensgebung mit sich. Im Großen und Ganzen waren die Protestanten früher nach Amerika gekommen und waren tendenziell erfolgreicher als die Katholiken. Die Unterschiede zwischen den zumeist katholischen *Patricks* und *Michaels* und den zumeist protestantischen *Williams* und *Roberts* spiegeln diese grundlegenden Bedingungen in der Heimat wider. Andere Namen, wie *John* und *James,* waren mehr oder weniger gleichmäßig auf die beiden Gruppen verteilt.

Da unter den in Philadelphia eingewanderten Iren mehr Katholiken als Protestanten waren und da es innerhalb des Namensbestandes der beiden Gruppen eine ähnliche Vielfalt (wenn auch nicht Übereinstimmung) von Namen gab, war zu erwarten, daß die Wahrscheinlichkeit, katholisch zu sein, bei Individuen mit häufig vorkommenden Namen etwas größer ist als bei Individuen mit selten vorkommenden Namen. *Tab. 7* stellt eine grobe Überprüfung dieser These dar. Hier ist der Prozentsatz der Männer mit besonders häufig vorkommenden Vornamen mit der

Tabelle 5
Iren 1806
Einzelne häufige Nachnamen nach Alter, Familienstand, Analphabetentum, Besitz und vertikalem Berufscode

	unter 30 %	verheiratet %	Analphabeten %	mit Grundbesitz %	durchschn. Grundbesitz p. Besitzer	mit Privateigentum %	durchschn. Privateigentum p. Besitzer	Anzahl
1. Kelly	35.4	57.9	8.6	8.4	$3554	32.8	$487	463
2. Smith	32.1	58.3	7.0	12.0	3212	32.9	422	374
3. Dougherty	32.8	54.3	8.9	8.0	1719	31.3	350	326
4. Murphy	33.8	60.2	8.0	8.3	4049	31.9	492	314
5. Campbell	29.8	60.1	5.2	12.5	4848	36.3	1776	248
6. McLaughlin	35.3	59.1	7.0	9.3	3862	40.5	630	215
7. Brown	33.8	60.2	7.5	12.4	4160	33.3	811	201
8. Wilson	39.7	61.8	3.5	11.6	3778	33.2	910	199
9. Gallagher	36.0	55.3	7.6	6.1	2342	34.5	316	197
10. O'Brien	33.3	53.9	7.2	6.1	3745	25.0	441	180
Grundgesamtheit	33.5	58.4	6.6	10.3	4824	32.7	836	37303

VERT

	0	1	2	3	4	5	6	7	8	9	Anzahl
1. Kelly	9.1	0.4	10.4	28.1	13.0	34.3	1.5	2.6	0.6	0.0	463
2. Smith	10.2	0.5	11.5	30.2	16.6	27.8	1.9	0.8	0.5	0.0	374
3. Dougherty	7.7	0.3	8.9	32.5	12.0	35.6	0.6	1.8	0.6	0.0	326
4. Murphy	8.6	1.0	7.3	31.8	12.7	35.0	1.9	1.6	0.0	0.0	314
5. Campbell	6.0	0.8	10.1	33.5	19.0	27.8	0.4	1.6	0.8	0.0	248
6. McLaughlin	9.3	0.9	10.7	30.7	13.0	32.6	0.5	1.4	0.9	0.0	215
7. Brown	4.5	1.5	15.9	32.3	17.4	23.4	2.5	1.0	1.5	0.0	201
8. Wilson	9.0	1.5	14.1	37.7	13.1	23.6	0.5	0.0	0.0	0.5	199
9. Gallagher	8.6	0.5	10.2	24.4	17.3	35.5	2.0	0.5	0.5	0.5	197
10. O'Brien	7.8	0.5	9.4	30.6	15.6	33.3	1.7	1.1	0.6	0.0	180
Grundgesamtheit	8.3	0.9	10.8	32.4	13.3	30.9	1.6	1.1	0.6	0.2	37303

Tabelle 6
Iren 1860
Einzelne häufige Vornamen nach Alter, Familienstand, Analphabetentum, Besitz und vertikalem Berufscode

	unter 30 %	verheiratet %	Analphabeten %	mit Grundbesitz	durchschn. Grundbesitz p. Besitzer	mit Privateigentum %	durchschn. Privateigentum p. Besitzer	Anzahl
J500 (John)	33.6	58.2	6.6	9.8	$4560	31.7	$784	6897
J700 (James)	33.6	58.4	6.4	11.2	4627	34.0	679	4782
W400 (William)	35.2	59.2	4.9	10.6	5147	33.4	986	3195
P700 (Patrick)	32.3	59.8	9.6	9.9	3960	30.8	679	2847
T500 (Thomas)	35.2	59.2	6.4	10.4	4253	32.5	783	2811
M200 (Michael)	34.5	59.1	8.3	7.5	3284	30.5	391	1879
R100 (Robert)	31.9	61.1	3.6	13.7	7713	38.4	1729	1249
D500 (Daniel)	32.4	59.0	8.8	10.2	3973	30.4	578	1134
E300 (Edward)	36.6	56.2	6.8	9.4	4214	30.4	633	873
Grundgesamtheit	33.5	58.4	6.6	10.3	4824	32.7	836	37303

VERT

	0	1	2	3	4	5	6	7	8	9	Anzahl
J500 (John)	8.3	0.8	10.7	32.9	13.2	31.0	1.5	1.1	0.5	0.2	6897
J700 (James)	7.6	0.8	11.0	34.1	13.0	30.2	1.5	1.0	0.7	0.1	4782
W400 (William)	8.1	1.5	11.0	36.9	12.8	25.7	2.0	1.1	0.6	0.3	3195
P700 (Patrick)	6.8	0.6	9.2	25.5	14.0	40.5	1.7	1.1	0.5	0.1	2847
T500 (Thomas)	6.6	0.7	11.8	31.5	13.6	32.7	1.6	1.0	0.5	0.1	2811
M200 (Michael)	7.7	0.3	9.7	29.1	12.3	37.0	2.1	1.2	0.6	0.0	1879
R100 (Robert)	7.8	1.6	11.5	38.9	12.5	23.7	1.4	1.0	1.2	0.3	1249
D500 (Daniel)	6.7	0.5	9.4	30.4	12.0	37.0	2.1	1.2	0.3	0.3	1134
E300 (Edward)	7.2	1.4	10.0	31.6	14.3	31.6	2.4	1.4	0.1	0.0	873
Grundgesamtheit	8.3	0.9	10.8	32.4	13.3	30.9	1.6	1.1	0.6	0.2	37303

Tabelle 7
Iren 1860
Häufigkeit des Nachnamens nach Vorname

	geringste Häufigkeit Code 0	Kategorien der Nachnamen (in %) Code 0-5	Code 6-11	größte Häufigkeit Code 11
J500 (John)	4.3	32.3	67.7	11.8
J700 (James)	4.0	32.9	67.1	10.2
W400 (William)	5.9	36.6	63.4	9.8
P700 (Patrick)	2.8	27.6	72.4	14.4
T500 (Thomas)	4.4	32.7	67.3	12.3
M200 (Michael)	3.3	30.4	69.6	12.8
R100 (Robert)	5.6	41.7	58.3	8.4
D500 (Daniel)	2.4	27.2	72.8	12.2
E300 (Edward)	3.4	27.0	73.0	12.0
Grundgesamtheit	4.2	32.9	67.1	11.3

Häufigkeit ihrer jeweiligen Nachnamen kreuztabuliert. Die weitgehend protestantischen *Williams* und *Roberts* hatten eher ungewöhnliche Nachnamen, die weitgehend katholischen *Patricks* hatten eher gewöhnliche Nachnamen. Folglich konnten die leichten Unterschiede zwischen häufig und selten vorkommenden Nachnamen sehr wohl durch subethnische Unterschiede innerhalb der eingewanderten Iren verursacht sein.

Um die von Stephan Thernstrom aufgeworfene Frage beantworten zu können, ob „die nicht Verfolgbaren" andere Mobilitätserfahrungen gemacht haben als die „Verfolgbaren", wurden die Lebensläufe von Iren mit häufig und mit selten vorkommenden Namen verglichen. Untersucht man Veränderungen bei Beruf und Besitz zwischen 1860 und 1870 (siehe *Tab. 8* und *9*)[23], so stellt sich heraus, daß Personen mit häufigen Vor- und Nachnamen in beiden Richtungen mobiler gewesen zu sein scheinen als Personen mit seltenen Namen. Diese Tendenz verschwindet aber, wenn der Abschneidepunkt heraufgesetzt wird, was vermuten läßt, daß die Unterschiede

23 Für die in Tab. 8 und 9 dargestellten Daten wurde die vertikale berufliche Mobilität grob wie folgt gemessen: VERT 1 und 2 wurde der Wert *1* zugeordnet; VERT 3 der Wert *2*, VERT 4, 5 und 6 der Wert *3* und VERT 0, 7, 8 und 9 der Wert *0*. War der neue Code einer Person in einem Jahr *0*, kam sie in die Kategorie *andere*. Andernfalls wurde ihre Mobilität als *abwärts*, *aufwärts* oder *unverändert* eingestuft. Änderungen beim Vermögen und Besitz wurden wie folgt berechnet: Der entsprechende Wert in Dollar (D) wurde mit der Formel $I = \ln(\frac{D+50}{50})$ in einen Index (I) umgewandelt. Der Dezimalbruch wurde gekürzt und der Wert für 1860 wurde von dem Wert für 1870 abgezogen. War das Ergebnis positiv, wurde die Mobilität als aufwärtsgerichtet angesehen, war es negativ als abwärtsgerichtet, und war es Null, dann als unverändert.

Tabelle 8
Iren 1860–1870
Nachnamen nach Mobilität

		Beruf				Grundbesitz			Privateigentum			
		abw.	unver-ändert	aufw.	sonst.	abw.	unver-ändert	aufw.	abw.	unver-ändert	aufw.	Anzahl
A. Total verkettete Datei (Gewichtung ≥14)												
geringste Häufigkeit	Code 0	13.4	53.0	13.4	20.1	4.5	70.1	25.4	22.4	32.8	44.8	134
	Code 0-5	12.7	57.4	12.7	17.1	5.5	71.7	22.8	24.8	38.0	37.3	1610
	Code 6-11	13.7	55.6	13.4	17.3	7.5	68.3	24.1	27.2	36.9	35.9	2845
größte Häufigkeit	Code 11	12.0	56.2	12.4	19.4	6.4	66.9	26.8	27.1	36.1	36.8	299
Grundgesamtheit		13.4	56.3	13.1	17.2	6.8	69.5	23.7	26.3	37.3	36.4	4455
B. Reduzierte Datei (Gewichtung ≥19)												
geringste Häufigkeit	Code 0	11.0	57.0	10.0	22.0	6.0	71.0	23.0	18.0	34.0	48.0	100
	Code 0-5	11.7	59.4	11.4	17.5	5.8	71.4	22.8	25.4	35.0	39.6	1135
	Code 6-11	11.2	59.3	12.0	17.5	7.6	67.4	25.0	28.8	35.8	35.4	1343
größte Häufigkeit	Code 11	6.7	53.8	18.5	21.0	5.0	69.7	25.2	23.5	41.2	35.3	119
Grundgesamtheit		11.4	59.4	11.7	17.5	6.8	69.2	24.0	27.2	35.4	37.3	2478

Tabelle 9
Iren 1860–1870
Vorname nach Mobilität

	Beruf				Grundbesitz			Privateigentum			
	abw.	unver-ändert	aufw.	sonst.	abw.	unver-ändert	aufw.	abw.	unver-ändert	aufw.	Anzahl
A. Total verkettete Datei (Gewichtung ≥14 als Abschneidep.)											
seltenere Namen	13.5	57.0	11.6	17.9	6.2	69.7	24.1	26.0	39.7	34.4	1752
häufigere Namen	13.2	55.8	14.1	16.8	7.2	69.4	23.4	26.6	35.8	37.7	2703
einzelne Namen											
J500 (John)	13.5	53.4	14.5	18.6	5.0	72.7	22.4	26.3	33.1	40.6	483
J700 (James)	10.6	57.9	14.4	17.1	7.4	64.8	27.8	28.9	32.4	38.7	432
W400 (William)	14.9	56.1	14.4	14.6	7.2	69.3	23.5	22.7	35.6	41.7	362
P700 (Patrick)	16.1	52.7	15.6	15.6	7.4	71.0	21.6	29.0	36.1	35.0	366
T500 (Thomas	10.4	57.0	15.9	16.8	9.1	65.2	25.6	24.7	38.1	37.2	328
M200 (Michael)	13.1	59.2	14.6	13.1	4.6	79.2	16.2	26.2	41.9	31.9	260
R100 (Robert)	13.6	54.3	9.2	22.8	9.8	62.0	28.3	28.3	32.1	39.7	184
D500 (Daniel)	13.1	56.2	12.4	18.3	8.5	68.6	22.9	25.5	40.5	34.0	153
E300 (Edward)	15.6	57.8	11.1	15.6	8.9	71.1	20.0	28.1	37.8	34.1	135
Grundgesamtheit	13.4	56.3	13.1	17.2	6.8	69.5	23.7	26.3	37.3	36.4	4455
B. Reduzierte Datei (Gewichtung ≥19 als Abschneidep.)											
seltenere Namen	12.0	59.5	10.7	17.7	6.7	68.6	24.7	27.2	36.1	36.7	1196
häufigere Namen	10.8	59.2	12.6	17.3	6.9	69.8	23.3	27.3	34.8	37.9	1282

einzelne Namen

J500	(John)	8.2	56.4	11.8	23.6	5.1	73.3	21.5	29.7	31.3	39.0	195
J700	(James)	8.0	62.8	13.8	15.4	7.4	67.6	25.0	29.8	27.7	42.6	188
W400	(William)	12.9	61.4	12.3	13.5	5.8	72.5	21.6	19.9	38.0	42.1	171
P700	(Patrick)	14.3	58.4	13.0	14.3	6.8	70.8	22.4	29.8	38.5	31.7	161
T500	(Thomas)	10.3	57.4	14.2	18.1	8.4	67.1	24.5	22.6	38.1	39.4	155
M200	(Michael)	11.5	61.2	15.1	12.2	5.8	77.7	16.5	27.3	40.3	32.4	139
R100	(Robert)	11.3	53.8	7.5	27.4	8.5	59.4	32.1	28.3	31.1	40.6	106
D500	(Daniel)	10.5	63.2	9.2	17.1	6.6	72.4	21.1	31.6	39.5	28.9	76
E300	(Edward)	12.1	58.2	14.3	15.4	8.8	62.6	28.6	29.7	30.8	39.6	91
Grundgesamtheit		11.4	59.4	11.7	17.5	6.8	69.2	24.0	27.2	35.4	37.3	2478

einfach auf den Einschluß einer größeren Anzahl falscher Paare bei den häufig vorkommenden als bei den seltenen Namen zurückgehen. Abgesehen von diesem Phänomen sind die Unterschiede sehr gering und scheinen kein Muster zu ergeben.

Die Untersuchung des *bias* der Namenshäufigkeit unter den in Irland geborenen Männern von 1860 deutet auf Abweichung auf der subethnischen Ebene und sollte die Aufmerksamkeit auf die sehr viel größeren Abweichungen in einer multiethnischen Bevölkerung, wie sie in den meisten amerikanischen Städten zu finden ist, lenken. In solchen Umgebungen kann die Datenverkettung dazu führen, daß kleinere ethnische Gruppen überrepräsentiert sind; deren Mitglieder sind leichter zu identifizieren, weil ihre Namen weniger häufig als die größerer ethnischer Gruppen sind. Deshalb werden bei der Verkettung vorteilhafter solche Dokumente wie die Manuskripte der Volkszählungen verwendet, da sie den jeweiligen Geburtsort explizit aufführen, und nicht solche Dokumente wie Adreßbücher, die diese kennzeichnende Angabe nicht enthalten. Im Zug der weiteren Forschung ist von uns die Erstellung eines Namensverzeichnisses auf der Grundlage von Dateien der Gesamtbevölkerung geplant, mit dem dann eine ethnische Aufgliederung der häufig vorkommenden Vor- und Nachnamen bereitstünde. Wenn dies für verschiedene Zeiten und Orte wiederholt wird, werden die Forscher das Ausmaß des ethnischen *bias*, den sie durch die Eliminierung bestimmter häufiger Namen aus ihren Verkettungsversuchen eingeführt haben, weit besser begrenzen können.

Schlußfolgerungen

Mit dem zunehmenden Interesse der Sozialhistoriker an der Rekonstruktion von individuellen Lebensläufen über die Zeit ist der Mangel an standardisierten und eigens für historische Daten entwickelten und verfeinerten Verfahren der Datenverkettung evident geworden. Das Hauptproblem ist hier, wie auch bei Untersuchungen zur sozialen Mobilität, unsere Unfähigkeit, uns auf eine gemeinsame Methode zu einigen. Forschungen, die unterschiedliche Verfahren der Datenverkettung und verschiedene Kategorien zur Klassifikation und Einordnung der Berufe verwenden, können nicht verglichen werden, sowenig wie die Raten der Beispiele für eine systematische Verwendung der Studien untereinander in unabhängige Variable umgewandelt werden können. Wie die Erbauer der Eisenbahnen im neunzehnten Jahrhundert die Vorteile der allgemein verwendeten Normalspurweite erkannten, so müssen sich auch die Historiker auf gemeinsame Konventionen einigen, wenn ihre Untersuchungen vergleichbar sein sollen. Wir haben deshalb unsere eigene Erfahrung mit der Datenverkettung als Erläuterung zu einer vielfach anwendbaren Methode dargestellt. Wir hoffen, daß unsere Erfahrungen mit der Datenverkettung der weiteren Diskussion und Zusammenarbeit als Vehikel dient mit dem Ziel, unter den Forschern, die mit der Verkettung von nominalen Daten aus historischen Quellen befaßt sind, zu einheitlicheren Verfahren zu gelangen.

Das historische Material über ein Individuum ist vom Standpunkt der Datenverkettung her keineswegs vollkommen. Das Hauptproblem dabei ist, daß die Merkmale der Individuen, ob sie nun in den Manuskripten der Volkszählung, in den Adreßbüchern oder anderen Quellen mitgeteilt werden, oft falsch aufgezeichnet sind. Wir fingen damit an, nach Möglichkeiten zu suchen, wie die lästigsten Probleme angegangen werden könnten. Besonders wichtig war die Schreibweise der Namen, denn zwei Varianten einer Schreibeweise, in denen der Forscher mühelos denselben Namen erkennen kann, können als solche vom Computer nicht identifiziert werden. Das *Russel-Soundex* war ein geeigneter Ausgangspunkt, aber zwei Verfeinerungsbereiche waren wichtig. Da wir mit Dateien arbeiteten, die auf der Transkription von handgeschriebenen Unterlagen aufbauten, war es zunächst nötig, nicht nur die Buchstaben zu berücksichtigen, die ähnlich klingen, sondern auch die, die ähnlich aussehen. Zweitens bestimmte unser Interesse an Personen, die Einwanderer waren, daß wir das *Soundex*-System modifizierten, um mit der besonderen Schreibweise fremdsprachiger Namen fertig zu werden. Wir glauben, daß unser *Soundex/Viewex*-System auch für andere Forscher der Sozialgeschichte nützlich sein kann, obwohl das Interesse an anderen ethnischen Gruppen bei jedem weiteren Anwendungsbereich eigene Modifikationen erfordert.

Zwar ist das Codieren der Namen entscheidend, aber doch kein Allheilmittel. Erstens kann kein Codierungssystem alle falschen Schreibweisen berücksichtigen und damit alle richtigen Verkettungen auffindbar machen. Außerdem ist die falsche Schreibweise nicht das einzige Problem, das bei der Datenverkettung auftaucht. Auch die Informationen über andere Variablen wie Alter und Geburtsort sind anfällig für Fehler. Es ist selbstverständlich wichtig, den Rahmen für die Vergleichsmerkmale bei der Verkettung unter den ausgeführten Restriktionen so weit wie möglich und zulässig zu stecken. So waren wir zum Beispiel durch die zusätzliche Aufnahme von Informationen über die Ehefrau eines Individuums (in vollem Bewußtsein des *bias*, den wir damit einführten) in der Lage, die Zahl der Verkettungen bei den verheirateten Männern zu verdoppeln. Am Ende kann aber doch kein maschinelles System der Datenverkettung – in der Tat auch kein einzelner Forscher, der eine Handbearbeitung vornimmt – mit absoluter Sicherheit entscheiden, daß jedes Datenpaar zutreffend verkettet ist. Es versteht sich deshalb von selbst, daß die Datenverkettung bei historischen Quellen probabilistischer Natur ist, und daraus ergibt sich die Notwendigkeit, ein System einzusetzen, das für die Handhabung von Wahrscheinlichkeiten entworfen ist, und nicht nur einfache *stimmt/stimmt nicht* Entscheidungen trifft.

Mit den Binärgewichtungen nähert man sich direkt der Vorstellung von der Wahrscheinlichkeit an, daß ein Datenpaar eine richtige Verkettung darstellt, nachdem eine Reihe von Merkmalen verglichen wurde. Gleichwohl haben wir versucht über eine bloße Feststellung des Systems der Binärgewichtung und seine mathematischen Formeln hinauszugehen, um damit zu zeigen, daß das System der Binärgewichtung selbst auf verschiedenen Stufen verbessert werden kann. Bei der Verwendung der Binärgewichtungen zum Beispiel muß die Entscheidung getroffen werden, ob sich die Entscheidung einfach darauf beziehen soll, daß eine Variable überein-

stimmt, oder darauf, daß diese Variable mit eigens festgesetzten, differenzierten Werten übereinstimmt. Das erste Verfahren ist, was die Rechenzeit anbelangt, billiger. Das zweite aber liefert genauere Indikatoren für die Konfidenz bei einer Verkettung. Bei der Beschäftigung mit Dateien, die aus zehntausenden von Fällen bestehen, war es für uns einfach nötig, nach spezifischen Werten bei den Variablen zu suchen. Die Wahrscheinlichkeit einer Verkettung ist größer, wenn ein selten vorkommender Name übereinstimmt, als wenn ein häufig vorkommender Name übereinstimmt, und die Wahrscheinlichkeit einer Verkettung ist größer, wenn von einem Individuum mitgeteilt wird, daß es nach einem Jahrzehnt zehn Jahre älter sei, als wenn das angegebene Altersintervall nicht zehn Jahre ist. Wir haben weiter festgestellt, daß die Verwendung von besonderen Gewichtungen für unübliche Werte der Variablen die Leistung unserer Verkettungsmethode erheblich verbessert hat, und zwar sowohl bei der Anzahl der vorgenommenen Verkettungen als auch bei der Konfidenz über ihre Zuverlässigkeit.

Darüber hinaus ist mit der Verwendung des Systems der Binärgewichtung eine effiziente Methode für die Auswahl der spezifischen Variablen gegeben, die in den Vergleich einbezogen werden sollen. Während es vorteilhaft ist, alle Merkmale zu verwenden, die nützlich erscheinen könnten, eliminiert das System der Binärgewichtung die Notwendigkeit und die Kosten, solche Merkmale zu vergleichen, die in hohem Maß miteinander korrelieren. In der Tat verringert die Einbeziehung von untereinander zusammenhängenden Variablen die Potenz der Binärgewichtungen sehr.

Wir haben auch darauf hingewiesen, daß Vorsicht vonnöten ist. Verkettete Dateien haben notwendig einen *bias*. Sie sind eindeutig nicht repräsentativ für alle Individuen, sondern enthalten nur die, die über einen Zeitraum hin am Ort geblieben sind. Unter diesem Aspekt ist der Zweck, dem eine verkettete Datei dienen soll, relevant, und man muß Untersuchungen zur Rate der Ortsfestigkeit strikt von Untersuchungen zur sozialen Mobilität trennen, um einen weiteren *bias* zu vermeiden. Der besondere Zweck, dem die verkettete Datei dienen soll, wird den Grad bestimmen, in dem inkorrekte Verkettungen bei der Analyse noch toleriert werden können. Es ist deshalb wichtig, daß jedes verkettete Paar einen Indikator für die Wahrscheinlichkeit erhält, daß es wirklich eine richtige Verkettung ist. Daß die Maßzahl der Binärgewichtung bei jedem Datenpaar mitaufgenommen wird, erfüllt diesen Zweck.

Aber die verketteten Dateien haben nicht nur darum einen *bias*, weil nur wirklich am Ort Verbliebene einbezogen werden. Der Verkettungsprozeß, den wir anwenden, bringt auch einen *bias* in die Datei im Hinblick auf Individuen mit ungewöhnlichen Merkmalen, besonders aber im Hinblick auf Individuen mit selten vorkommenden Namen. Dennoch hat es uns die Verwendung von zusätzlichen Merkmalen, wie sie in den Manuskripten der Volkszählung verfügbar sind, ermöglicht, die Wirkung eines solchen *bias* zu bestimmen. Insbesondere waren wir in der Lage den Grad der Abweichung einer Person mit selten vorkommenden von einer Person mit häufig vorkommendem Namen zu bestimmen. Daß Unterschiede zwischen solchen Individuen bestehen, ist eindeutig, aber unsere Schlußfolgerungen lauten, daß

solche Unterschiede normalerweise nicht genügend ausgeprägt sind, um großes Interesse erwecken zu können.

Es kann nicht geleugnet werden, daß man bei der Verkettung von historischen Daten von Fallgruben umgeben ist. Doch ist auch der mögliche Erfolg bei der Datenverkettung entsprechend hoch. Unsere eigene Forschung erbringt bereits solche ersten Früchte. In diesem Aufsatz lag die Betonung auf den methodologischen Problemen, aber es sollen doch noch kurz einige unserer vorläufigen Befunde zusammengefaßt werden.

Erstens: unsere bis heute einzige, nicht auf männliche Erwachsene beschränkte Untersuchung beinhaltete die Verkettung aller Schwarzen, der Männer, Frauen und Kinder in dem Jahrzehnt von 1850 bis 1860. Obwohl wir bei diesem Experiment nur solche Personen verketteten, deren Nachnamen mit dem Buchstaben *T* anfingen, ähnelte diese Gruppe in jeder Hinsicht allen Schwarzen. Ihr gehören ungefähr ein Viertel der erwachsenen Männer und Frauen und, höchst ermutigend, ein Drittel der Kinder, männlichen und weiblichen Geschlechts gleichermaßen, an[24]. Zweitens: bei unserem einzigen Versuch, Stichproben zu verketten, konnten wir ungefähr ein Fünftel der im Land geborenen männlichen Weißen in der Dekade von 1850 bis 1860 wiederfinden (nachdem wir den Stichprobenfaktor von eins zu sechs verbesserten)[25]. Drittens: da wir in dieser Dekade einen wesentlich niedrigeren Anteil an irischen und deutschen Einwanderern verketteten – 14 % resp. 13 % – bestätigte sich unser Verdacht, daß in einem Einreisehafen die Ortsfestigkeitsraten für Einwanderer niedriger als für die schwarze oder weiße Eingeborenengruppe sind. Mit anderen Worten: viele Iren und Deutsche, die in unseren Querschnittsnetzen hängenblieben, hatten nicht die Absicht, in Philadelphia zu bleiben. Als letztes: Die Raten für die Ortsfestigkeit von zehn, zwanzig und dreißig Jahren für erwachsene, männliche Iren und Deutsche sind bedeutend niedriger als die in allen anderen Untersuchungen angegebenen Raten[26].

Läßt sich zum jetzigen Zeitpunkt etwas Definitives über die Ortsfestigkeitsraten im Amerika des neunzehnten Jahrhunderts sagen? Alle früheren Untersuchungen geben an, daß nach Ablauf von zehn Jahren etwa 40 bis 60 Prozent der Bevölkerung wiedergefunden werden konnten[27]. Da von der Grundstruktur unseres Datenver-

24 Alle unten aufgeführten Ortsfestigkeitsraten sind als vorläufig anzusehen. Sie sind in dreifacher Hinsicht bedingt: (1) die Datenverkettung wurde nur mit den Variablen ausgeführt, die unsere Mobilitätsstudie nicht beeinflussen würden; (2) die Raten berücksichtigten die Sterblichkeit nicht; (3) die Raten berücksichtigen Verkettungsfehler nicht.

25 Wir verfolgten 1957 im Lande geborene, männliche Weiße über 18 Jahre, deren Nachname mit dem Buchstaben *S* oder *B* begann – eine eins zu sechs Stichprobe aller entsprechenden Männer im Jahr 1850 – zu einer analogen Stichprobe von im Lande geborenen, männlichen Weißen 1860. Hätte die Ortsfestigkeit bei 100 % gelegen, hätten wir ein Sechstel der Gruppe von 1850 zu finden erwartet oder 321 Männer. In Wirklichkeit aber fanden wir ein Fünftel dieser Zahl.

26 Siehe Tab. 9:1: *Persistence rates in selected urban communities, 1800–1968. . .* , in: Stephan Thernstrom, *The Other Bostonians*, S. 222.

27 Ibid.

kettungsprogrammes her erwartet werden konnte, daß es die Zahl der vorgenommenen Verkettungen erhöhen würde, hätte man annehmen können, daß wir das in anderen Studien vorhandene Ausmaß der Unterbewertung schätzen könnten. Aber wie wir gesehen haben, mußte eine solche Unterbewertung nicht einmal geschehen sein. Wir finden, um es noch einmal zu sagen, einen sehr viel niedrigeren Anteil von Verkettungen. Bei der Studie, die der unseren am nächsten steht, hat Stuart Blumin mit Adreßbüchern 38 % der erwachsenen männlichen Einwohner von Philadelphia verkettet (1850–1860). Wie oben erwähnt, fanden wir weniger als 20 % der Gesamtbevölkerung der Stadt (die vier ethnischen Gruppen zusammengenommen). Zum Teil läßt sich dieser Unterschied mit der Vollständigkeit der verwendeten Dokumente erklären. Adreßbücher weisen wahrscheinlich einen *bias* im Hinblick auf beständigere Individuen auf. Aber wahrscheinlicher ist doch, daß das Verfahren, das bei der Adreßbuchverkettung zur Identifikation verwendet wurde – Übereinstimmung des Namens – dafür verantwortlich ist. Es ist durchaus möglich, daß Blumin und andere, die Adreßbücher verwenden, eine erhebliche Anzahl Personen verketten, die nach unseren Maßstäben nicht verkettet würden. Wenn dies häufig genug geschah, kann die Zahl der Verkettungen, die nicht hätten gemacht werden dürfen, aber doch gemacht wurden, der Zahl der Verkettungen, die gemacht worden wären, aber nach unserem Verfahren nicht gemacht wurden, entsprechen. Die Fehler würden sich in diesem Fall gegenseitig aufheben und die Ortsfestigkeitsrate könnte dieselbe bleiben. Solange aber nicht alle Forscher ihre Methoden explizit darstellen, sodaß sie ausgewertet und verglichen werden können, oder noch besser, nach standardisierten Datenverkettungsmethoden verfahren, ist es unmöglich, definitive Aussagen über die Rate der Ortsfestigkeit zu machen. Obgleich die Probleme, denen wir gegenüberstehen, deutlich machen, daß niemals alle wirklich Ortsfesten verkettet werden, auch nicht mit den Informationen in den Manuskripten der Volkszählung und einem Computer zu unseren Diensten, wird es bei weiterer strenger Forschung möglich werden, die Ortsfestigkeitsraten exakt zu schätzen. Bis dahin aber wenigstens bleibt die Rate der Ortsfestigkeit der Bevölkerung im Amerika des neunzehnten Jahrhunderts eine offene Frage.

Summary: Record Linkage

The Philadelphia Social History Project (PSHP) at the University of Pennsylvania has developed a computerized "record linkage" program. The essay identifies several of the problems involved in the conceptualization of record linkage and summarizes the methodological and statistical approach used in the implementation of our automated procedures.

Record linkage may be defined as the bringing together of information derived from independent sources concerning a particular historical entity be it a person, family, institution or event. Record linkage is a methodological procedure central to much of the "new" research in urban and social history taking plase in the U.S.

Although record linkage as a generic term encompasses a great many document types and cross-time ("horizontal") and over-time ("vertical") linkages, substantive discussion here focuses on the latter category and outlines the procedures used in the linkage of decennial population census records for all Irish and German immigrants to Philadelphia in the years 1850–1880. In this essay, emphasis is placed on the distinction between the linkage of records and their subsequent analysis. Variables used in the identification procedure will bias the linked files; researchers are cautioned to indicate, preferably on each case, which variables were used in the linkage procedures.

The record linkage approach used at the PSHP is best described as "probabilistic" for comparisons of an "either-or" nature are eschewed. Three distinct conceptual steps in the record linkage process are described: comparison space, comparison function and decision rule. The comparison space is derived from the two files which are to be linked and consists of all pairs obtained by drawing one individual from each initial file. The comparison function is the manner by which the members of each pair are compared. The decision rule takes the outcome of the comparison step and designates a pair as linked, non-linked or as indeterminate.

Included in the discussion are: the Russell "soundex" coding system to overcome spelling errors and a variant of this procedure, "viewex", to overcome errors attributable to illegibility; a "binit" weighting system in which each linkage variable is separately considered to determine conditional probabilities; and the specification of a "cut-off" point to identify records which are thought to be accurately linked.